The Anâtaxis phylogenetic method. 1. The algorithm : building from a dissimilarity matrix a phyletic tree accounting for homoplasy and lineage-dependent heterogeneity of transformation rates

Autor(en): **Bittar, Gabriel**

Objekttyp: **Article**

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

Band (Jahr): **55 (2002)**

Heft 1

PDF erstellt am: **01.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-740287

Archs Sci. Genève	Vol. 55	Fasc. 1	pp. 1-8	Juin 2002

THE ANÂTAXIS PHYLOGENETIC METHOD. 1. THE ALGORITHM - BUILDING FROM A DISSIMILARITY MATRIX A PHYLETIC TREE ACCOUNTING FOR HOMOPLASY AND LINEAGE-DEPENDENT HETEROGENEITY OF TRANSFORMATION RATES

BY

Gabriel BITTAR

(Ms. reçu le 3.1.2002, accepté le 23.5.2002)

ABSTRACT

The Anâtaxis phylogenetic method. 1. The algorithm - building from a dissimilarity matrix a phyletic tree accounting for homoplasy and lineage-dependent heterogeneity of transformation rates. - Anâtaxis is a phylogenetic method using as input the semi-matrix of dissimilarities (between terminal taxa) and a known outgroup. It is a splitting method reconstructing the evolutionary tree by working up from the root to the terminal nodes, detecting on the fly homoplasy and allowing for heterogeneity of transformation rates between the lineages - as such it is not a clustering, neighbour-joining method. It aims to be both phyletically correct and operationally rapid.

Key-words: Cladistic Maximum Parsimony, Dissimilarity matrix, Evolutionary tree, Heterogeneity of transformation rates, Homoplasy, Numerical Taxonomy Phenetics, Outgroup-based method, Phylogenetic methods, Splitting method.

There is a need for a phylogenetic method of analysis (WILLS, 1994) capable of producing rapidly, for a large amount of data, a reasonably correct dendrogram, even in the presence of homoplasy and heterogeneity of evolutionary rates between the different lineages. It should operate on a direct analysis of the dissimilarities between all pairs of terminal nodes of the tree (quantitative data which are metric but not additive distances, because of multiple hits per site/character and *homoplasy*), as some numerical taxonomic phenetics (NTP) methods do (SOKAL, 1986), rather than on the matrix of the states of characters, as cladistic maximum parsimony (CMP) methods do (HENNIG, 1966; SWOFFORD *et al.*, 1996). The reason is that there is too much information within a matrix of the states of characters when the aim is simply to produce a phyletic tree: practically, CMP methods are very slow, and fundamentally they are not consistent, i.e. there is not necessarily convergence of outputs when the data on a given problem grow. On the other hand, the semi-matrix of dissimilarities between pairs of terminal nodes, with its much-reduced information content, carries for all practical means all the necessary information for the accomplishment of this task. The trick is to avoid the phyletic pitfalls associated with a clustering-like, neighbour-joining type of analysis, i.e. the unaccounting

[1]Universities of Geneva and Lausanne, po box 281, American River, Kangaroo Island, South Australia 5221, bittar@isb-sib.ch

of homoplasy, the joining together of terminal nodes having evolved within slowly-transforming lineages, and the pushing out to a basal (out-group) position of those terminal nodes having evolved within quickly-transforming lineages.

The Anâtaxis method aims to address this phylogenetic problem without resorting to a probabilistic approach (FELSENSTEIN, 2001), with its ingrained slowness. With this method it is not intended to optimalise any kind of scalar measure, such as minimalising the length of the global tree: Anâtaxis is a splitting method which only tries to define a tree that is compatible with the dissimilarity matrix, and that minimalises *ad hoc* evolutionary hypotheses, by working up from the root to the terminal nodes - as such it is not a clustering, neighbour-joining method. It can be applied to any kind of data obtained from (self-)reproducing, transformable objects, i.e. related, evolutive objects, with the sole conditions that one (at least) of these evolving objects can be defined as outgroup to the others, and that one can construct a symmetrical matrix of object-to-object dissimilarities. We present here the main steps of the Anâtaxis algorithm.

1. Choose a set of **s+o** homologous sequences (i.e. sequences for which a common evolutionary ancestor is the cause for similarity) within which two main clades may be defined, the smaller clade of cardinal o, with known internal structure, being the outgroup to the larger, ingroup clade, of cardinal s and of which the internal structure is to be defined.

2. From the s+o aligned homologous sequences, calculate the (s+o)(s+o-1)/2 values of the semi-matrix Δ of pairwise dissimilarity values, or (s+o) x (s+o) symmetrical matrix with all members of the main diagonal having value zero.

3. Correct each Δ dissimilarity value for sequencing errors and affect it with a "fuzziness" function pertaining to standard error (due to paucity of dissimilar sites) and to polymorphism / uncertainty (Bittar 1998, unpublished). Matrix D is thus created.

In other words, 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 D_{oi}, constituting e.g. the first line of the whole dissimilarity semi-matrix.

In the following, simple example of a dissimilarity matrix (a much more complete example is also provided, BITTAR, 2002), clade 6 constitutes the out-group o of cardinal 1, the s taxa 1 to 5 constitute the in-group I of cardinal 5:

sequence	1	2	3	4	5	
6	D_{61}	D_{62}	D_{63}	D_{64}	D_{65}	D_{oi} OUT-IN vector
5	D_{51}	D_{52}	D_{53}	D_{54}		
4	D_{41}	D_{42}	D_{43}			
3	D_{31}	D_{32}				D_{ij} IN-IN sub-matrix
2	D_{21}					

Better, there can be two or more starting out-groups, producing as many different dissimilarity matrices from which one produces an arithmetic (or algebraic) mean OUT-IN vector D_{oi} (where the index i designates a member of the in-group I);

e.g., if there are two basal out-groups, o_1 and o_2, that we use together as outgroup **o+**, such as we have at the root of the tree an unresolved trichotomy (o_1, o_2, I), we can write

$D_{oi}^{+} = (D_{io_2} + D_{io_1})/2$.

Even better, each starting out-group can be a whole clade with known internal structure, again allowing for the calculation of a weighted $\mathbf{D_{oi}}$ 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 can write

$\mathbf{D}_{io} = (2D_{io_3} + D_{io_2} + D_{io_1})/4$.

4. Median-normalise on the weighted outgroup all the original dissimilarity values.

All the D_{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. Empirically, using the median of the D_{oi} for normalising gives good results. Hence, all the difference values between this median and the Δ_{oi} are calculated :

$diff_{oi} = med(D_{oi}) - D_{oi}$

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

$D_{ij}^{*} = D_{ij} + diff_{oi} + diff_{oj}$.

Example: let us consider a 5-taxa ingroup of which we want to know the internal phyletic structure, with taxon 6 (or 6+) being used as the starting, normalising outgroup-taxon.

Dissimilarity matrix of Di_j , of which $med(D_{oi}) = 9$ (we write in bold the corresponding $diff_{oi}$ and $diff_{oj}$):

sequence		1	2	3	4	5	
o = 6(+)	**diff**	**-1**	**0**	**+3**	**0**	**0**	
6(+)	**0**	10	9	6	9	9	D_{oi} OUT-IN vector
5	**0**	9	8	5	4		
4	**0**	9	8	5			
3	**+3**	4	3				D_{ij} IN-IN sub-matrix
2	**0**	3					

Normalised dissimilarity matrix of D_{ij}^* :

sequence		1	2	3	4	5	
o = 6(+)	*diff*	*-1*	*0*	*+3*	*0*	*0*	
6(+)	*0*	9	9	9	9	9	D_{oi}^* OUT-IN vector
5	*0*	8	8	8	4		
4	*0*	8	8	8			
3	*+3*	6	6				D_{ij}^* IN-IN sub-matrix
2	*0*	2					

5. For each triad {a,b,c} of normalised dissimilarities D_{ba}^*, D_{ca}^* and D_{cb}^* thus formed within the ingroup, define the pair of equality ($\approx$) / inequality ($<<$) relations according to the "fuzziness" conditions, i.e. is an inequality ($<$) clear enough so as to be treated as such (symbolised by $<<$), or does it deserve to be treated practically as an equality (symbolised by $\approx$) ?

There are 4 possible pairs of ($\approx$ or $<<$) dissimilarities relations :

$D_{ba}^* \approx D_{ca}^* \approx D_{cb}^*$ (three practically identical normalised dissimilarities),
$D_{ba}^* << D_{ca}^* \approx D_{cb}^*$ (one small, two big and practically identical, normalised dissimilarities),
$D_{ba}^* \approx D_{ca}^* << D_{cb}^*$ (two small and practically identical, one big, normalised dissimilarities),
$D_{ba}^* << D_{ca}^* << D_{cb}^*$ (three clearly different normalised dissimilarities).

6. For each triad from within the normalised ingroup, chose the most likely (most parsimonious in evolutionary terms) trichotomous solution(s) - if there are two possible solutions, apply weighting to them or a resolution procedure.

For s taxa, the number of triads that can be formed is the number of possible combinations of 3 - distinct - elements among s elements, i.e.

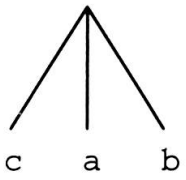
$$C^3_s = s!/[(s-3)!\ 3!] = s(s-1)(s-2)/6 .$$

Dissimilarities to trichotomies, correspondence table

OUT-IN : $D_{oc}^* = D_{ob}^* = D_{oa}^*$ (o = outgroup)

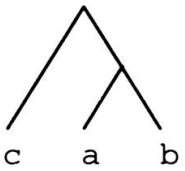
'z homopl. x' = z partially homoplasic with x
the sign '+' indicates a relatively rapid transformation rate within a lineage, from the time of divergence at the root of the outgroup and ingroup

- $D_{ba}^* \approx D_{ca}^* \approx D_{cb}^*$



The unresolved trichotomy case.

- $D_{ba}^* \ll D_{ca}^* \approx D_{cb}^*$



The ultrametric-like resolved trichotomy case.

- $D_{ba}^* \ll D_{ca}^* \ll D_{cb}^*$

[a homopl. c]
or [b+ homopl. o]

The homoplasic resolved trichotomy.

[Note the possible ambiguity about the precise homoplasy relationship; for the partial homoplasy to imply the normalising outgroup o, it is necessary for the '[b+ homopl. o]' alternative to appear in *each* triad involving b; this is a stronger constraint than for the '[a homopl. c]', which requires to be true in each triad involving *both* a and c].

- $D_{ba}^* \approx D_{ca}^* \ll D_{cb}^*$

[a homopl. c]
or [b+ homopl. o]

OR

[a homopl. b]
or [c+ homopl. o]

The homoplasic *and* phyletically ambiguous case, which requires weighting or a resolution procedure.

If $D_{ba}^* = D_{ca}^* \ll D_{cb}^*$, the philetic ambiguity is unavoidable and might need more data to be resolved;

If $D_{ba}^* < D_{ca}^* \ll D_{cb}^*$, the philetic ambiguity can be avoided by deciding to interpret the first inequality as $D_{ba}^* \ll D_{ca}^* \ll D_{cb}^*$.

[As for the possible ambiguity about the precise homoplasy relationship in each of the two alternative triads, refer to the note in the preceding case].

In our example, there are 10 triads that can be formed between the 5 taxa / sequences (1 to 5), each leading to a trichotomy.

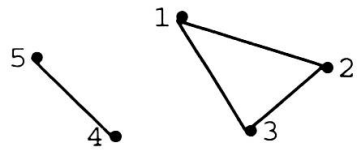
For triad {5,2,1}, the three normalised dissimilarities are $2 << 8 = 8$, we are in an ultrametric-like situation thus the phyletic relationship between these three taxa is the homoplasy-free resolved trichotomy (5,(2,1)).

We successively treat in the same manner the nine other triads, in each case we notice an ultrametric-like ($D_{ba}^* \ll D_{ca}^* \approx D_{cb}^*$) phyletic relationship between the three taxa involved , thus all triads lead to a resolved trichotomy :

(5,(2,1)), (5,(3,1)), (5,(3,2)), (4,(2,1)), (4,(3,1)), (4,(3,2)), (3,(5,4)), (2,(5,4)), (1,(5,4)), (3,(2,1)).

7. We create a space composed of s points. For each triplet of points (taxa) where there is a (c,(b,a)) solution, we create a link between b and a, thus in a perfectly dichotomous global tree we end up with a partition of two and only two subsets.

In our example, we get the following final partition :



The links between taxa 5 and 4, and between taxa 2 and 1, are triple, the ones between taxa 3 and 2, and between taxa 3 and 1, are double.

From now on, subsets {5,4} and {3,2,1} may each be considered as outgroup to the other. If the clade d = {5,4} (obviously there's no need here to resolve its internal structure) is considered as outgroup to the clade {3,2,1}, we can calculate, with $D_{di} = (1/2)[D_{5i} + D_{4i}]$, a new, smaller matrix of dissimilarities :

taxon/seq.	1	2	3	
d	D_{d1}	D_{d2}	D_{d3}	new D_{oi} OUT-IN vector
3	D_{31}	D_{32}		new D_{ij} IN-IN sub-matrix
2	D_{21}			

In our example, this gives, with med(D_{oi}) = 8 :

sequence		1	2	3	
o = d	**diff**	**-1**	**0**	**+3**	
d = {5;4}	**0**	9	8	5	D_{oi} OUT-IN vector
3	**+3**	4	3		D_{ij} IN-IN sub-matrix
2	**0**	3			

Go back to step 4.

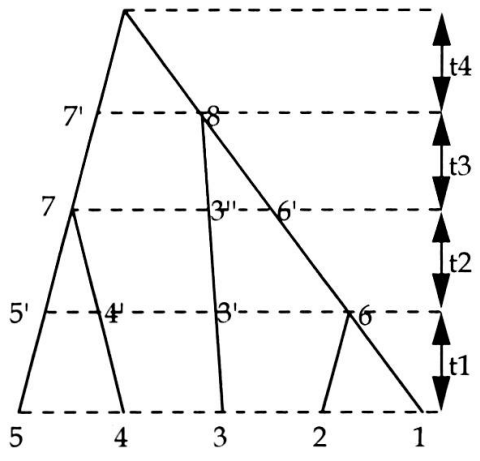
It is important to note that, in this *second iteration*, the normalisation is *not* performed on already normalised values, but on the original values, because any normalisation is an approximation procedure for taking into account the phenetic biasing produced by the heterogeneity of lineage-specific evolution rates, and it is better to avoid approximating on approximations : so there is no such thing as a D_{ij}**.

So, in our example, this gives for the new normalised matrix :

sequence		1	2	3	
o = d	*diff*	*-1*	*0*	*+3*	
d = {5;4}	*0*	8	8	8	D_{oi}^{*} OUT-IN vector
3	*+3*	6	6		D_{ij}^{*} IN-IN sub-matrix
2	*0*	2			

For triad {3,2,1}, the three normalised dissimilarities are 2 << 6 = 6, we are in an ultrametric-like situation thus the phyletic relationship between these three taxa is the homoplasy-free resolved trichotomy (3,(2,1)).

We can now draw the final tree, which is perfectly bifurcating :



LA MÉTHODE PHYLOGÉNÉTIQUE ANÂTAXIS. 1. L'ALGORITHME - RECONSTITUTION À PARTIR DE LA MATRICE DE DISSIMILITUDES D'UN ARBRE PHYLÉTIQUE TENANT COMPTE DE L'HOMOPLASIE ET DE L'HÉTÉROGÉNÉITÉ INTER-LIGNÉES DES VITESSES DE TRANSFORMATION

Anâtaxis est une méthode phylogénétique utilisant comme input la semi-matrice de dissimilitudes (entre taxons terminaux) et un outgroup connu. Un arbre évolutif est reconstitué par divisions successives, en opérant depuis la racine jusqu'aux nœuds terminaux, détectant en cours de route les cas d'homoplasie et tenant compte de l'hétérogénéité entre les lignées des vitesses de transformation - ce n'est donc pas une méthode de 'clustering' ou de 'neighbour-joining'. Son objectif est d'être à la fois phylétiquement correcte et rapide dans son fonctionnement.

REFERENCES

BITTAR, G. 2002. The Anâtaxis phylogenetic method. 2. An example - reconstituting a whole dendrogram. *Arch. Sc.* 55: 9-20.

FELSENSTEIN, J. 2001. *PHYLIP (Phylogeny Inference Package)*, v. 3.6, Seattle Wash., distr. by the author at http://evolution.genetics.washington.edu/phylip.html, Dep. of Genetics, University of Washington.

HENNIG, W. 1966. *Phylogenetic Systematics*. Univ. Illinois Press, Urbana.

SOKAL, R. R. 1986. Phenetic taxonomy: theory and methods. *Annu. Rev. Ecol. Syst.* 17: 423-442.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL & D. M. HILLIS. 1996. Phylogenetic Inference, pp. 407-514. *In*: Molecular Systematics, 2nd Edition (D. M. Hillis, C. Moritz & B. K. Mable, eds.). *Sinauer Associates, Sunderland, MA*.

WILLS, Ch. 1994. Phylogenetic Analysis and Molecular Evolution, pp. 175-201. *In*: Biocomputing: Informatics and Genome Projects (D. W. Smith ed.). *Academic Press*.