The Anâtaxis phylogenetic method. II. Inferring a tree taking into account : homoplasy and heterogeneity of evolutionary rate over phyletic lineages

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

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: **22.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-740279

Archs Sci. Genève	Vol. 50	Fasc. 2	pp. 169-175	Septembre 1997

THE ANÂTAXIS PHYLOGENETIC METHOD. II. INFERRING A TREE TAKING INTO ACCOUNT HOMOPLASY AND HETEROGENEITY OF EVOLUTIONARY RATE OVER PHYLETIC LINEAGES

BY

Gabriel BITTAR

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, has been 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 and the last step of this method have been presented in the preceding paper. Here, the fourth step in the Anâtaxis procedure is described. It consists in looking the ensemble of proposed triadic trees, for a sub-outgroup which may be the starting point of a new iteration of steps two to four, on a smaller matrix no longer containing the original outgroup ("peeling" process). In case no such sub-outgroup can be easily determined, pairing of 2p taxa is performed, on the basis of the normalised dissimilarity, and steps two to four are begun again, but with p less taxa in the ingroup sub-matrix. Other path alternatives exist in case of difficult, noisy input data. Once the dissimilarity matrix has been entirely "peeled", the fifth and last step consists in reconstituting the global tree, in different manners according to the way that noise has been taken into account. Hence, the basic idea in the Anâtaxis global tree reconstruction approach is to find the tree which fits best to the topological frequencies "spectrum" obtained from a given data-set, by comparing this "spectrum" to a reference tree-spectra table, rather than to try, classically, to optimalise a numerical function. This procedure, which is in accordance with the view that the numerical data composed by the dissimilarity matrix ought to be treated as a "fuzzy" set of values, is technically difficult, but the positive result of this up-to-down reconstitution approach is to strongly reduce the probability of artefacts in the global tree due to homoplasy and evolutionary rate heterogeneity between lineages.

RECONSTRUCTING THE GLOBAL TREE : ANÂTAXIS FOURTH STEP

Basic principle

If Anâtaxis detects that a taxon e was external in *all* asymmetrical triads (resolved trichotomies) where it appeared (let us say it is taxon 4), it considers that this taxon e constitutes a new outgroup (a sub-outgroup) for the new, smaller, ingroup (taxa 1 to 3 in our example), and the whole triads analysis process is renewed, but on a smaller ingroup matrix of dissimilarity (this is called the "peeling" procedure) :

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

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@expasy2.hcuge.ch

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

All the original components Δ_{oi} of the newly defined OUT-IN vector (where index $_o$ stands for sub-outgroup taxon 4, and index $_i$ designates the ingroup taxa 1 to 3) are made identical to their median value, and the newly normalised $\Delta_{ij}^{\star}$ are analysed according to the triads correspondence table (if the clustering method has been chosen, it is necessary to cluster the components of the new $\overline{\Delta}_i^{\star}$ perfectly ordered vector). It is important to note that, in this second run, 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 bias 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 $\Delta_{ij}^{\star\star}$.

The great advantage of such a "peeling" procedure, is that this up-to-down reconstitution approach strongly reduces the probability of artefacts in the global tree due to homoplasy (and evolutionary rate heterogeneity) between lineages : cladistically the nearer the reference outgroup, the lesser the probability of homoplasy of i or j with o, and also, but this is less important because it is already addressed during the normalisation procedure, of a strong difference of evolutionary rates heterogeneity between i and j. This is an important advantage on "neighbour-joining", down-to-up phenetical procedures.

Complex sub-outgroup : plurality of leaves

Now, what happens when there is no unique taxon e, i.e. if sub-outgroup e is constituted by a whole clade of taxa ? In fact, for any set of s distinct homologous sequences (s leaves, or terminal taxa), e could be constituted from 1 taxon up to s/2 taxa – if s is an even number; up to (s-1)/2 taxa otherwise –. In this case of multiple taxa per sub-outgroup, the method consists in looking for the taxa which appear the most often as external in resolved trichotomies, and comparing these less than 100% occurrence cases with the numbers from a **tree-spectra table.**

Indeed, in accordance with our view that the numerical data composed by the dissimilarity matrix ought to be treated as a "fuzzy" set of values, our basic idea in the Anâtaxis global tree reconstruction approach is to find the tree which fits best to the topological frequencies "spectrum" obtained from a given data-set, by comparing this "spectrum" to a reference tree-spectra table, rather than to try, classically, to optimalise a numerical function. Let us see exactly how.

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.

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

For any member of an outgroup clade composed of o taxa, the relative number of times this taxon would appear as the external branch (externality frequency) in any trichotomous tree containing this taxon and two other taxa from the remaining population of $s-1$ taxa, can be expressed by the following formula :

$$f_{external/outgroup} = [C^2_{(s-o)} + \{0 \text{ to } C^2_{(o-1)}\}] / C^2_{(s-1)}$$
$$= [(s-o)(s-o-1) + \{o \text{ to } (o-1)(o-2)\}] / (s-1)(s-2)$$

The reason for a minimum and a maximum within this formula is that if the outgroup is composed of more than 2 taxa, it may be a (partially) resolved polychotomy and thus have an internal cladistic structure, i.e. a taxon within this outgroup could, for example, itself be an outgroup to the o-1 other taxa of this group (in which case the maximum in the formula would apply for this member of the outgroup).

Let us say that, for $s = 13$ taxa, there is a taxon e_1 that was external in 69.7% of all triads in which it appeared; according to the externality frequency table, this implies that, very likely, there is an outgroup composed of $o = 3$ taxa, in which this taxon e_1 is itself cladistically external to the two other taxa e_2 and e_3 of this outgroup : if this is indeed the case, there should be 2 other taxa (e_2 and e_3) affected both with externality frequencies of 68.2%, and the internal structure of the outgroup is $(e_1,(e_2,e_3))$. If there are 3 taxa affected with frequencies 68.2%, then the internal structure of the outgroup is (e_1,e_2,e_3), i.e. an unresolved trichotomy.

The task of defining a sub-outgroup from the externality frequencies remains tractable only if all the highest frequencies are strictly for members of this sub-outgroup. Therefore, it is necessary that

$$\min\{f_{external/outgroup}\} > \max\{f_{external/ingroup}\} .$$

Since the externality frequency for a member of the ingroup is

$$f_{external/ingroup} = [C^2o + \{0 \text{ to } C^2_{(s-o-1)}\}] / C^2_{(s-1)}$$
$$= [o(o-1) + \{0 \text{ to } (s-o-1)(s-o-2)\}] / (s-1)(s-2) .$$

it is necessary that

$$(s-o)(s-o-1) > o(o-1) + (s-o-1)(s-o-2)$$
$$\Leftrightarrow \quad s > (o^2 + o + 2)/2 .$$

As can be seen from the tree-spectra table at the end of this paper, if the outgroup is composed of two taxa ($o = 2$) this implies that the total number of sequences (terminal taxa) should be greater than four ($s > 4$); if $o = 3$: $s > 7$; if $o = 4$: $s > 11$; if $o = 5$: $s > 16$; if $o = 6$: $s > 22$; and so on.

If $\min\{f_{external/outgroup}\}$ is strictly greater than $\max\{f_{external/ingroup}\}$, it is possible to define which taxa are members of the sub-outgroup simply from a comparison of the externality frequencies — found by the Anâtaxis triads analysis — with those displayed in the table, but only in clear cases where all these frequencies effectively belong to the interval $[\min\{f_{external/outgroup}\} ; \max\{f_{external/outgroup}\}]$. If $\min\{f_{external/outgroup}\}$ is not strictly greater than $\max\{f_{external/ingroup}\}$, there is great difficulty in defining which taxa are members of the sub-outgroup simply from the externality frequencies, and more information, obtainable from the Anâtaxis triads analysis, is necessary for proceeding further.

Complex sub-outgroup : internal structure

There is another difficulty when the size of the potential sub-outgroup grows, which lies not only in defining which taxa are members of it, but also in defining its internal cladistic structure. Let us consider this with the still simple case, where the outgroup is composed of $o = 4$ taxa, e_1, e_2, e_3 and e_4, which implies, for $s > 11$, that the four taxa with highest externality frequency are the four members of the outgroup. For $s = 13$ taxa, these four taxa are affected with externality frequencies ranging from 54.5% to 59.1%.

The internal structure of this clade of 4 taxa can still easily be inferred from the distribution of externality frequencies for taxa e_1, e_2, e_3 and e_4, because there is only (whatever the value of s) one intermediary frequency between the minimum and the maximum : an e_h ($h = 1, 2, 3$ or 4) may be external in 3/3, 1/3 or 0/3 triads composed of e_h and two other members of the tetradic outgroup. Thus, if the externality frequencies distribution is e_1 : 59.1%, e_2 : 56.1%, and twice 54.5%, then the outgroup structure is the asymmetrical perfectly resolved tetrachotomy $(e_1, (e_2, (e_3, e_4)))$. If the distribution is e_1 : 59.1%, and thrice 54.5%, then the outgroup structure is $(e_1, (e_2, e_3, e_4))$. If the distribution is 56.1% for e_1 and e_2, and 54.5% for e_3 and e_4, then the outgroup structure is $(e_1, e_2, (e_3, e_4))$. If the distribution is four times 54.5%, then the outgroup structure is an unresolved tetrachotomy, (e_1, e_2, e_3, e_4). If the distribution is four times 56.1%, then the outgroup structure is a symmetrical resolved tetrachotomy; in this case, it must be noted that the distribution of frequencies is not sufficient to precisely resolve the structure, and other information is necessary, namely the pairing associations : if the Anâtaxis analysis detects that e_1 and e_2 are always paired together, and that e_3 and e_4 are always paired together, then the internal structure of the outgroup is $((e_1, e_2), (e_3, e_4))$.

Clearly, when the outgroup is composed of more than 4 taxa, i.e. when $o > 4$, it is much less easy to infer its internal structure directly from the distribution of externality frequencies for taxa e_1, e_2, e_3, e_4, e_5, etc., simply because of the complex diversity of pentachotomous and higher-order poly(cho)tomous trees, and the correlated complexity of externality frequencies distributions (moreover, one must not forget the increasing noise in the data...).

Yet, it is important that the internal cladistic structure of a sub-outgroup be known, because, the Anâtaxis procedure will consider this whole clade as one taxon, composed from all its constitutive taxa, each with its own weighting. This has implications for the calculation of mean Δ_{oi} OUT-IN dissimilarity values, and also for the calculation of the corresponding uncertainty values.

Methods for bypassing difficulties in the global tree reconstruction

This is why, *if*, during the "peeling" procedure of the dissimilarity semi-matrix, there is no possibility of easily defining the members of a sub-outgroup, and of determining its internal cladistic structure, another important feature of Anâtaxis is, rather classically, its ability to pair all the p pairs of taxa that were always paired together within the (resolved) trichotomous trees, thus creating with each pair a new taxon of which all the dissimilarities are simply the means of the original dissimilarities (and the uncertainties are redefined as explained in the preceding paper). Remember that Anâtaxis operates with this pairing (neighbour-joining) procedure on the basis of normalised dissimilarity values, so this is a legitimate move : the risk of joining together taxa simply because their lineages have evolved slowly has been drastically reduced. This step being done, Anâtaxis begins again the trichotomy correspondence procedure, with the original outgroup o unchanged but with p less taxa within the ingroup.

We must insist on the fact that even if the pairing procedure has been made legitimate by the normalisation process, in case the program detects both a clear sub-outgroup clade with a clear cladistic structure, *and* pairing possibilities within the ingroup, the pairings within the ingroup are not immediately implemented. The reason for this is that the next run, using, after "peeling" the old outgroup, the new, nearer outgroup, shall allow a better estimation of the heterogeneity of evolutionary rates among the ingroup lineages, accordingly a better normalisation of the dissimilarity values of the new, smaller IN-IN semi-matrix, and therefore a better detection of pairing possibilities within the ingroup.

Now, in a case where the data is so inconclusive or ambiguous that neither the "peeling" nor the pairing procedure is possible, the user may :

– opt for a majority-rule consensus, with a threshold of his choice (larger than 50%, but smaller than the 100% of the strict consensus), either for defining the potential members of sub-outgroup e, or for the pairing process;

– and/or look for a consensus (strict or, if really necessary, majority-rule) only on the triads devoid of homoplasy (on the basis that these triadic solutions may be the most important source of chaos in the phyletic reconstitution).

A DYNAMIC PROCESS

It must be evident by now that if the Anâtaxis method is rather straightforward, not making use of very sophisticated mathematical tools, on the other hand it is not, because of the complexities we have seen, a method that is easy to automatize and thus to implement in a computer program. Consequently, even if a user may a *priori* define a series of potential choices that would have to be done for an entire procedure, it is better for him to stand by the computer, look at the output of each intermediate step of the "peeling" procedure, and then properly choose what the next step should be. Since the actual calculations are really rapid, this is not too handicapping, but it must be clear that the Anâtaxis program, at this stage of development, still requires much work and attention on the part of interested users.

Whatever, before concluding, let us be reminded that the whole Anâtaxis procedure can be performed either on only the original dissimilarity matrix; or this matrix plus the two matrices defining the upper-boundary and lower-boundary dissimilarity values; or this matrix plus the noise-generated matrices, each producing its own tree, a consensus tree being then obtained from all these trees.

And finally, it must be underlined that Anâtaxis shall allow for a dynamic treatment of data input relatively to the addition of new "leaves". In presence of a data set with no missing values, if we wish to add to the initial ingroup a new terminal taxon z, it is only necessary to add to the existing Δ_{ij} semi-matrix the new vector Δ_{zj}, which can be done in a straightforward manner. Once we have proceeded with normalisation, we look directly for the taxa which are most neighbour to z. Once the plausible clade to which z seems to belong has been defined, z is not joined to the members of this clade, but renormalisation is proceeded with, on the basis of the next outgroup according to the old tree (i.e. the tree obtained without z); and so on. In this way, it is not mandatory to recalculate *de novo* a tree when the data set is expanded. This potentiality is a very useful feature to avoid the quadratic growth of calculation time with the number s of "leaves", and has become indispensable in an age where taxa diversity in genes data-bases augments explosively for some popular genes (such as the chloroplastic *rbc*L).

CONCLUSION

In summary, Anâtaxis is a dissimilarity-matrix and outgroup-based, triadic trees-compatibility phylogenetic program, which, by taking into account both the temporal heterogeneity of the rates of evolution among different branches, and the possibility of homoplasy, has the advantage of avoiding the systematic biases common to distances methods, while keeping their inherent rapidity. Together with TakâmolE, which allows for missing data even if there is spatial heterogeneity of substitution rates, they offer a new approach for phylogenetic inference that has proven, on different sets of molecular sequences data, to be efficient and useful for evolutionary biologists.

Tree-spectra table

s = total number of sequences ("leaves" or terminal taxa).
o = number of taxa that are members of the outgroup.
o = 1 to s/2 if s is even, (s-1)/2 if s is uneven.

Number of times, relative to all triads where it appears,
that a taxon that belongs to an outgroup clade is the external member of the triad.
The lowest possible frequency is given first : [(s-o)(s-o-1)]/[(s-1)(s-2)],
then the highest possible frequency : [(s-o)(s-o-1)+(o-1)(o-2)]/[(s-1)(s-2)].
Then the lowest and highest possible frequencies
among the remaining s-o taxa of the ingroup are also given :
[o(o-1)]/[(s-1)(s-2)] and
[o(o-1)+(s-o-1)(s-o-2)]/[(s-1)(s-2)]

	o = 2	3	4	5	6	7	8	9	10
s = 4	33.33%								
	33.33%								
	33.33%								
	33.33%								
5	50.00%								
	50.00%								
	16.67%								
	33.33%								
6	60.00%	30.00%							
	60.00%	40.00%							
	10.00%	30.00%							
	40.00%	40.00%							
7	66.67%	40.00%							
	66.67%	46.67%							
	6.67%	20.00%							
	46.67%	40.00%							
8	71.43%	47.62%	28.57%						
	71.43%	52.38%	42.86%						
	4.76%	14.29%	28.57%						
	52.38%	42.86%	42.86%						
9	75.00%	53.57%	35.71%						
	75.00%	57.14%	46.43%						
	3.57%	10.71%	21.43%						
	57.14%	46.43%	42.86%						
10	77.78%	58.33%	41.67%	27.78%					
	77.78%	61.11%	50.00%	44.44%					
	2.78%	8.33%	16.67%	27.78%					
	61.11%	50.00%	44.44%	44.44%					
11	80.00%	62.22%	46.67%	33.33%					
	80.00%	64.44%	53.33%	46.67%					
	2.22%	6.67%	13.33%	22.22%					
	64.44%	53.33%	46.67%	44.44%					
12	81.82%	65.45%	50.91%	38.18%	27.27%				
	81.82%	67.27%	56.36%	49.09%	45.45%				
	1.82%	5.45%	10.91%	18.18%	27.27%				
	67.27%	56.36%	49.09%	45.45%	45.45%				
13	83.33%	68.18%	54.55%	42.42%	31.82%				
	83.33%	69.70%	59.09%	51.52%	46.97%				
	1.52%	4.55%	9.09%	15.15%	22.73%				
	69.70%	59.09%	51.52%	46.97%	45.45%				
14	84.62%	70.51%	57.69%	46.15%	35.90%	26.92%			
	84.62%	71.79%	61.54%	53.85%	48.72%	46.15%			
	1.28%	3.85%	7.69%	12.82%	19.23%	26.92%			
	71.79%	61.54%	53.85%	48.72%	46.15%	46.15%			
15	85.71%	72.53%	60.44%	49.45%	39.56%	30.77%			
	85.71%	73.63%	63.74%	56.04%	50.55%	47.25%			
	1.10%	3.30%	6.59%	10.99%	16.48%	23.08%			
	73.63%	63.74%	56.04%	50.55%	47.25%	46.15%			
16	86.67%	74.29%	62.86%	52.38%	42.86%	34.29%	26.67%		
	86.67%	75.24%	65.71%	58.10%	52.38%	48.57%	46.67%		
	0.95%	2.86%	5.71%	9.52%	14.29%	20.00%	26.67%		
	75.24%	65.71%	58.10%	52.38%	48.57%	46.67%	46.67%		
17	87.50%	75.83%	65.00%	55.00%	45.83%	37.50%	30.00%		
	87.50%	76.67%	67.50%	60.00%	54.17%	50.00%	47.50%		
	0.83%	2.50%	5.00%	8.33%	12.50%	17.50%	23.33%		
	76.67%	67.50%	60.00%	54.17%	50.00%	47.50%	46.67%		
18	88.24%	77.21%	66.91%	57.35%	48.53%	40.44%	33.09%	26.47%	
	88.24%	77.94%	69.12%	61.76%	55.88%	51.47%	48.53%	47.06%	
	0.74%	2.21%	4.41%	7.35%	11.03%	15.44%	20.59%	26.47%	
	77.94%	69.12%	61.76%	55.88%	51.47%	48.53%	47.06%	47.06%	
19	88.89%	78.43%	68.63%	59.48%	50.98%	43.14%	35.95%	29.41%	
	88.89%	79.08%	70.59%	63.40%	57.52%	52.94%	49.67%	47.71%	
	0.65%	1.96%	3.92%	6.54%	9.80%	13.73%	18.30%	23.53%	
	79.08%	70.59%	63.40%	57.52%	52.94%	49.67%	47.71%	47.06%	
20	89.47%	79.53%	70.18%	61.40%	53.22%	45.61%	38.60%	32.16%	26.32%
	89.47%	80.12%	71.93%	64.91%	59.06%	54.39%	50.88%	48.54%	47.37%
	0.58%	1.75%	3.51%	5.85%	8.77%	12.28%	16.37%	21.05%	26.32%
	80.12%	71.93%	64.91%	59.06%	54.39%	50.88%	48.54%	47.37%	47.37%
21	90.00%	80.53%	71.58%	63.16%	55.26%	47.89%	41.05%	34.74%	28.95%
	90.00%	81.05%	73.16%	66.32%	60.53%	55.79%	52.11%	49.47%	47.89%
	0.53%	1.58%	3.16%	5.26%	7.89%	11.05%	14.74%	18.95%	23.68%
	81.05%	73.16%	66.32%	60.53%	55.79%	52.11%	49.47%	47.89%	47.37%