

Phylogenetic analysis of Gregory of Nazianzus' Homily 27

Anne-Catherine Lantin¹, Philippe V. Baret¹, Caroline Macé²

¹Université catholique de Louvain – AGRO – GENA – 1348 Louvain-la-Neuve – Belgique

²Katholieke Universiteit Leuven – HIW – 3000 Leuven – België

lantin@gena.ucl.ac.be

Abstract

For a few years, biology and philology have come together by sharing a tool, namely the phylogenetic methods. Initially developed to establish relations between living organisms, these methods can also help reconstitute the history of multiple copies of an ancient text, which accumulated variations as time went by. This was already materialized by a few fruitful collaborations between philologists and biologists.

In this paper, the phylogenetic method of maximum parsimony was applied to 105 manuscripts containing the text of Gregory of Nazianzus' Homily 27. We aimed to compare results obtained by this phylogenetic analysis with what is known by philologists about the history of the text. The tree obtained agrees with philological evidences, but some undetermination remains for a few manuscripts.

This experiment brings new support to the pertinence of applying phylogeny approaches to manuscript traditions, which constitute a valid and powerful tool of classification.

Keywords: phylogeny, stemmatology, classification of manuscripts

1. Introduction

Phylogeny refers to the evolutive history of biological species, hypothetically reconstituted and represented as a tree. Stemmatology, a discipline from the field of philology, also builds trees, or stemmata, that tell how an unique ancient text gave birth to multiple different copies, as a result of variations introduced by scribes in the course of time. Both disciplines have thus in common that they try to infer the evolutionary history of their objects.

1.1. Methodological and conceptual similarities

In both disciplines, the scientific investigation can be seen as a three-step process. First, observing and comparing objects (living organisms or manuscripts) in order to build a dataset. Second, classifying the objects into groups according to their similarities. Third, if the data required is available, finding their origin and inferring their evolutive history.

Objects can be analyzed only if they are reduced to a certain amount of characteristics. For living organisms, these characteristics consist in morphological, physiological, or molecular features (sequences of genes and proteins) and are called *characters*. For manuscripts, these features consist in all the positions within the text which vary from one reading to another; they are called *variant readings*.

Transmission and *variation* are key concepts in evolution, for organisms as well as for manuscripts.

Transmission of characters follows the basis hypothesis that an object inherits all its characteristics from the same ancestor: this is known as the unique ascendance hypothesis. However, this assumption is sometimes violated. In philology, it occurs when one manuscript is copied from two different templates, and this phenomenon is called *contamination*. In biology, it occurs when two species cross to give birth to a new one (*hybridation*) or when genetic material is transferred from one species to another (this phenomenon is known as *horizontal gene transfer*).

Characters are transmitted from one object to its descendants, but they also change in the course of time. In biology, *variation* is produced by genetic events (mutations); in philology, by alterations of the texts, introduced intentionally or unintentionally by those who copied them.

1.2. Phylogenetic methods

Phylogenetic methods are computerized algorithms that create trees, which graphically show relationships between objects. Different methods can be used and debate continues to center around the merits of competing methods of phylogeny inference. They can be applied to any set of data which contains objects subject to evolution (temporal changes), and as a result, they are also adapted to the construction of 'trees of manuscripts' (stemmata).

Extension of biological algorithms to philological applications was carried on over several manuscript traditions, giving consistent results (Mooney *et al.*, 2003 ; Howe *et al.*, 2001 ; Barbrook *et al.*, 1998 ; Robinson and O'Hara, 1993).

The objective of the present paper is to use biological approaches to build a stemma of a Homily written by Saint Gregory of Nazianzus (ca. AD 330-390), one of the Cappadocian Church Fathers¹ and a bishop of Constantinople. The stemma aims to represent the evolution of Gregory's text from the 5th century (when the text was translated into Latin) to 1550 (when first printed versions of the text started to circulate).

2. Material and methods

2.1. Dataset

The dataset consists in a matrix of 141 lines and 556 columns. Lines correspond to all the extant manuscripts that contain Gregory of Nazianzus' Homily 27, and columns indicate all the positions within the text where the reading is different from one manuscript to another. As more than two variant readings can be found for the same position, the coding is not binary. For instance, entries in a matrix column can go from 0 to 7 when there are eight different variant readings for the same position. These codes do not reflect any chronological order, but '0' means that the manuscript has the same reading as the base text, an arbitrarily chosen manuscript used as a fixed comparison point. In our case, the base text is an edition from the 18th century, not included in the 141 manuscripts we studied.

The 141 lines correspond to 135 Greek manuscripts, 1 printed edition (EA), 1 artificial witness (Tanc) and 4 translations. Two translations are in Syriac (Tsyrl and Tsyrl2), one is in Latin (Tlat), and one is in Armenian (Tarm). We left aside 36 of the Greek manuscripts: 25 manuscripts are uninformative for editing the text, 8 manuscripts are too incomplete, and no microfilm is available for 3 manuscripts.

¹ The title of "Church Father" is given to the most important Greek and Latin authors from the 2nd to the 6th century, whose writings are considered founding texts of the tradition of the Church.

For these 105 witnesses, 86 out of 556 characters are constant and 282 variable characters are parsimony-uninformative, which means that they do not bring any information to build a phylogenetic tree under the parsimony principle. As a result, only 188 characters were taken into account for the analysis.

2.2. Phylogenetic analysis

The phylogenetic analysis was performed using PAUP* version 4.0b10 (Swofford, 1998). It is the result of a four-step process.

2.2.1. Choosing the optimality criterion of parsimony

Phylogenetic methods search for the tree that meets an optimality criterion by evaluating all possible individual trees generated out of the data. The optimality criterion called parsimony is based on the assumption that the most likely tree is the one that requires the fewest number of changes to explain the data. The basic premise of parsimony is that objects sharing a common characteristic do so because they inherited that characteristic from a common ancestor (in our case, that manuscripts sharing the same variant reading do so because they were copied from the same template). However, conflicts with that assumption often occur, because sharing a common characteristic may be due to independent parallel evolutionary events. For example, two scribes may independently introduce the same mistake in their copies. In biology, these independent evolutionary changes are gathered together under the term *homoplasy*. Homoplasies are regarded as ‘extra’ steps or hypotheses that are required to explain the data. Parsimony operates by selecting the tree or trees that minimize the number of evolutionary steps, including homoplasies. The number of steps required to fit the data with a given tree is referred to as the score of the tree.

Other optimality criteria exist, such as Maximum Likelihood, but they require precise assumption of the mechanisms of change and precise parameters of evolution; no such model exists for the evolution of manuscripts.

2.2.2. Conducting a heuristic search

When the number of manuscripts is small, it is possible to evaluate each of the possible trees. This exhaustive search guarantees finding the best tree (the most parsimonious one). But the number of possible trees rapidly increases with the number of manuscripts and an exhaustive search becomes impractical even for the most powerful computers (for ten manuscripts, there are more than 34 millions possible unrooted trees). When the number of trees is large, a heuristic strategy is used. A heuristic approach is essentially an algorithm in which an initial tree is selected, then rearrangements are sought to improve the tree. There are numerous algorithms of rearrangements. The one used in this experiment is the TBR search (tree bisection-reconnection): a provisional tree is cut into two subtrees at a branch, and these two subtrees are then reconnected by joining two branches, one from each subtree, to generate a different tree. This is tried for all possible pairs of branches of the two subtrees to generate many different trees.

2.2.3. Rooting the trees

Rooting a tree means choosing a point on the unrooted tree as representing the earliest time in the evolutionary history of the objects. An unrooted tree only represents groupings between the objects; a rooted tree shows a sequence of events, starting from the root. Rooting a tree can be done by designating a point in the middle of a branch as the common origin, or by selecting any one of the objects as an outgroup (an object designated as an outsider to the

other objects). In outgroup rooting, the point where the outgroup meets the ingroup is the ancestor of the whole ingroup. In this experiment, we rooted all the equally parsimonious trees with "Tanc". This witness was chosen as an outgroup because it is an artificially made manuscript, built with variant readings from Tarm and Tlat considered to be more ancient than the others.

2.2.4. *Computing the consensus tree*

Maximum parsimony methods often produce several parsimonious trees. In our analysis, more than 400 000 trees with the same score were generated. To make sense of this huge amount of equivalent trees, we built a consensus tree which sums up the common patterns observed in all the most parsimonious trees. In a strict consensus tree, any conflicting branching patterns for a set of objects among some rival trees are not resolved and result in a multifurcating branching pattern. This is why in Figure 1, some manuscripts are not organized in clusters but all have the same status according to the vertical branch they originate from.

3. Results

3.1. Tree of the complete tradition

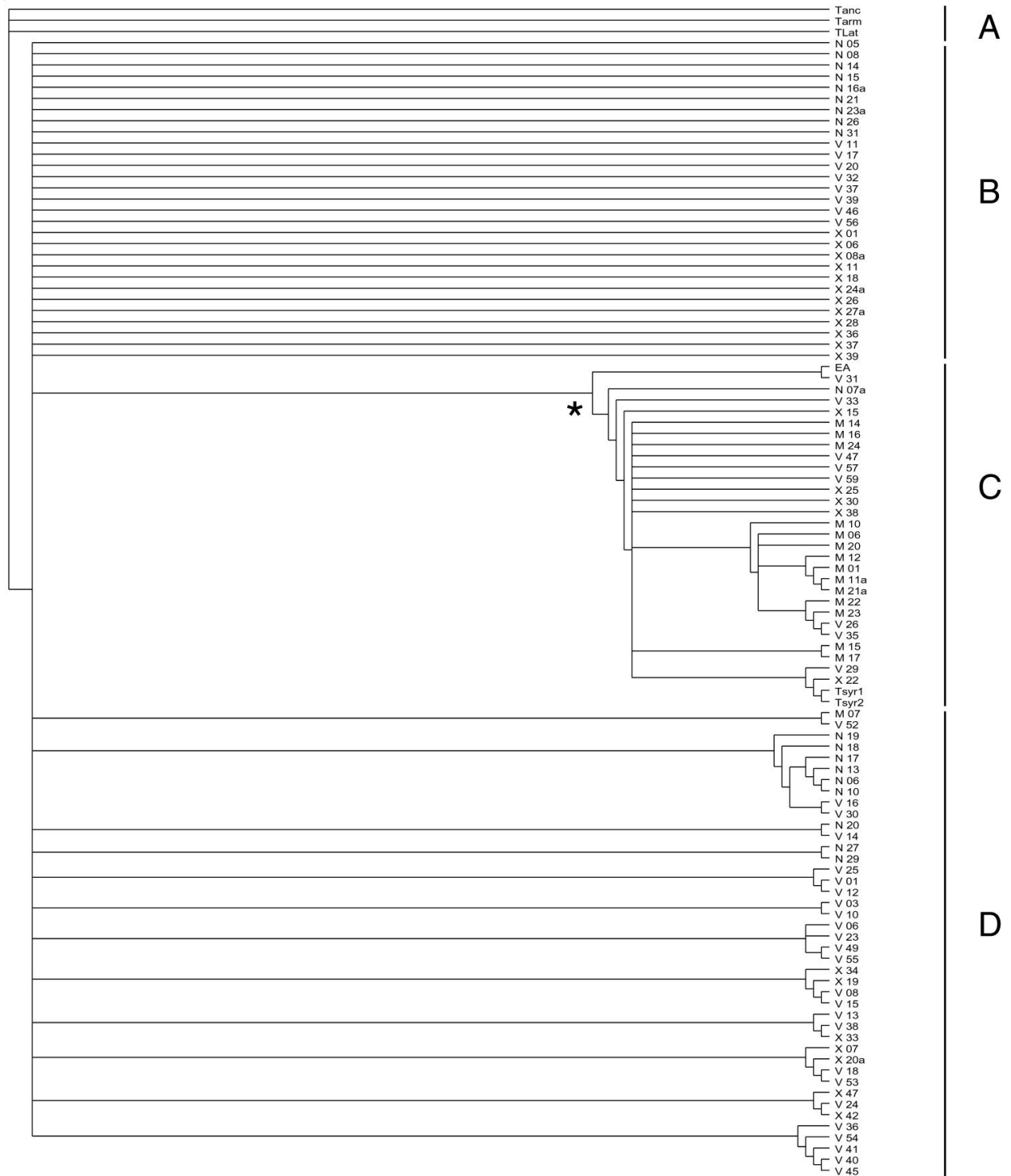


Figure 1. Consensus tree of the complete manuscripts tradition, created by PAUP* and visualized with TREEVIEW. Codes on the right refer to the main clusters.

3.2. Interpreting the tree

3.2.1. The point of view of the biologist

The tree is rooted with Tanc, which clusters with two translations, Tlat and Tarm [A].

Three other groups can be defined:

Group [B] contains 29 manuscripts, all deriving from the same vertical branch. This does not mean that all the manuscripts were copied from the same template, but rather that the relationships between these manuscripts are undefined, due to a lack of discriminant information.

Group [C] contains 31 clearly differentiated manuscripts. Within this group, clear substructures such as [Cm] and [Cs] are well defined (Figure 2).

Group [D] contains 42 manuscripts, clustered in small independent groups of 2 to 8 manuscripts. All manuscripts from each group should share a common ancestor, but no higher structure is observed.

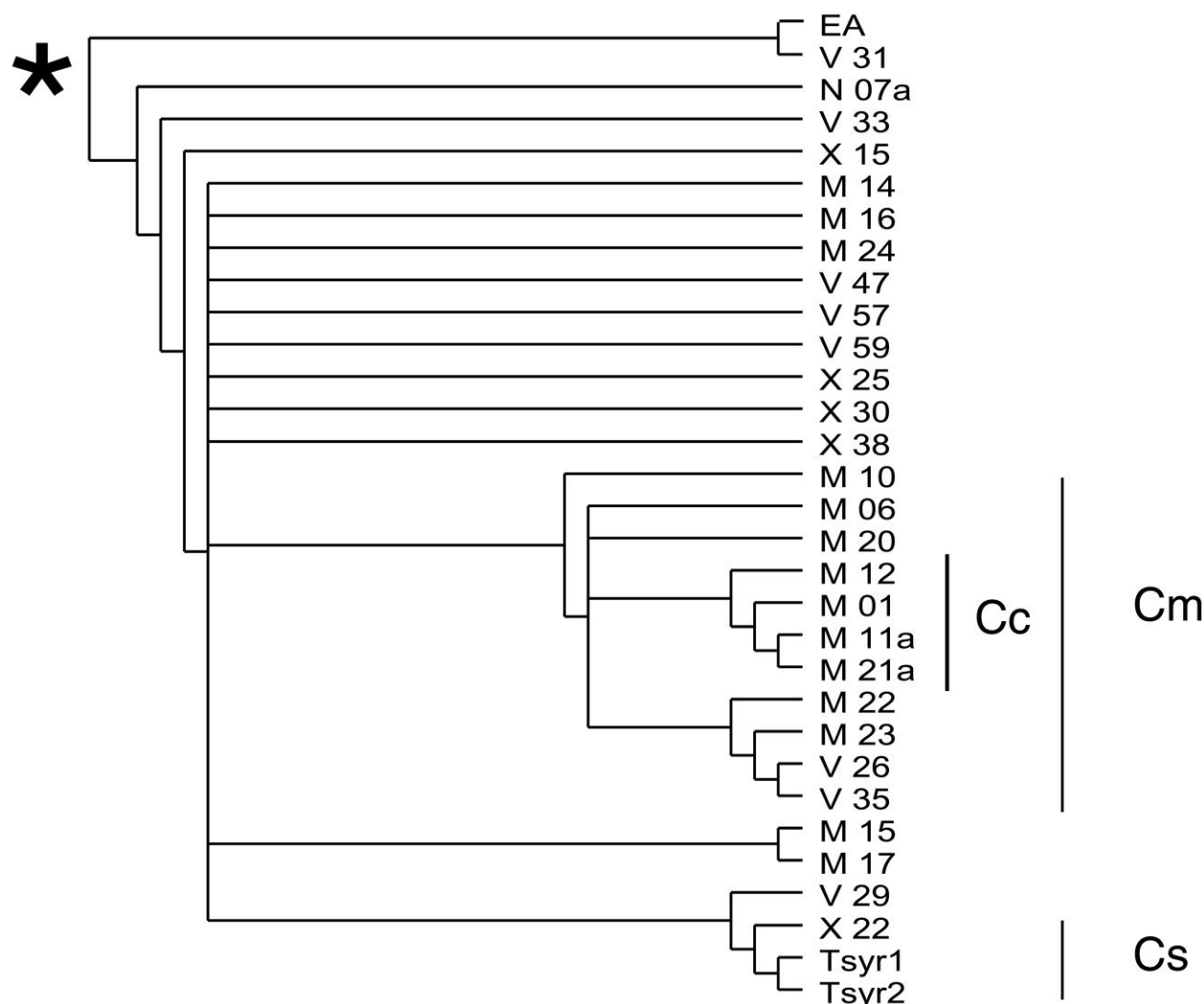


Figure 2. Details of the consensus tree of figure 1: the C group.

3.2.2. The point of view of the philologist

Reading the tree in Figure 1, the philologist can draw some conclusions about each one of the groups.

Group [A] clusters Tarm and Tlat, close to the root Tanc. As Tanc is a consensus made from Tarm and Tlat, it is not surprising to find them together in group [A]. It is also interesting to note that the two Syriac translations do not appear in this group, even if in Tsy1 and Tsy2 most data are missing, exactly like in Tarm and Tlat. The presence of missing data is due to the way the translations were processed and discrepancies coded between Greek standard text and translations. All other statistical or computerized methods that were applied precedently (Baret, 2004; Macé, 2003) were unable to correctly handle missing data: Tsy1 and Tsy2 were always misclassified and misplaced, next to Tarm and Tlat, with which they have in common only missing data.

Manuscripts from group [B] are indeed difficult to classify, and this undetermination represents fairly well the problems faced by philologists with these manuscripts, due to the absence of discriminant variant readings.

Most of the manuscripts in the [Cm] group are from Southern Italian origin, as opposed to the other manuscripts. Interestingly enough, the Syriac translations (Tsy1 and Tsy2) [Cs] belong to the same main group [C] as this Southern Italian sub-group [Cm]. It is indeed well known that strong cultural connections existed between Syria and Southern Italy in the Byzantine world before the time when our oldest preserved Greek manuscripts were written. It is thus plausible that Tsy1 and Tsy2 were translated from the same template as the one used for the Greek copies from the [Cm] group. Paleographical evidence also shows that all manuscripts from group [Cc] were copied by Calabrian monks. It is thus exact to group them in the same cluster.

Two subgroups from group [D] are also corroborated by external philological criteria. The first group contains manuscripts N.18, N.17, N.13, N.06, N.10, V.16 and V.30, which share the same order of homilies and the second group contains manuscripts V.36, V.54, V.41, V.40 and V.45, which share the same 'mise en page' and marginal notes.

4. Conclusion

By applying the maximum parsimony method to the manuscripts containing Homily 27, we generated a tree which is validated by philological evidence and which offers new hypotheses that could be investigated by philologists. The most parsimonious tree could not be found, due to limits in computing time, so the tree presented is a consensus of more than 400 000 equally parsimonious trees. In this tree, the two Syriac translations were correctly handled, which was not the case when other statistical or computerized methods were used.

From a technical point of view, phylogenetic methods can be easily transferred to philological data. The tree of Figure 1 is the first complete tree ever drawn showing relationships between all manuscripts of an Homily of Gregory of Nazianus, although research on this complicated manuscript tradition has been carried on for more than one hundred years. Some of the groups are corroborated by external evidences (same origin, similarities of scripture or 'mise en page') and despite the fact that some relationships remain undetermined, the structure of the tree broadly agrees with what is known about the text.

A conceptual analysis reveals that some problems of phylogenetic inference in biology also apply in philology. For instance, *hybridation* and *horizontal gene transfer* correspond to the philological phenomenon of *contamination*; *homoplasies* also exist for manuscripts; and *rooting the trees* requires in biology as well as in philology some information over the age of manuscripts or the chronology of variant readings. *Contamination* and *homoplasies* were not

taken into account in this analysis, but identifying ambiguous manuscripts and characters and removing them from the dataset could help reduce the number of equally parsimonious trees.

Although phylogenetic inference still has to face technical and conceptual challenges, we are convinced that it is the most powerful tool offered to the philologist in order to explore a manuscript tradition.

PAUP procedure

```
execute nazianze-complet.nex
taxset exclus = 2-26 31 39 75 86 87 133 28 116 126 71 64
delete exclus / only
set criterion = parsimony
charset noninform = uninformat
exclude noninform / only
hsearch addseq = simple reftax = 65
>Select 'Reset Maxtrees'
>Select 'automatically increase by 200'
outgroup Tanc
roottrees
savetrees file = nazianze.tre
contree/ treefile = consensus.tre
```

Acknowledgements

We would like to thank all the members of the Center for the Study of Gregory of Nazianzus (<http://nazianzos.fltr.ucl.ac.be>), and particularly Marc Dubuisson who built all the databases. We would also like to thank A. Vanderpoorten for his helpful comments on the computer analysis.

References

- Barbrook A.C. *et al.* (1998). The phylogeny of the *Canterbury Tales*. *Nature*, vol. (394): 839.
- Baret P.V. *et al.* (accepted). Experimental phylogenetic analysis of a Greek manuscript tradition. In *Proceedings of the Washington Academy of Sciences*.
- Hall B.G. (2001). *Phylogenetic Trees Made Easy*. Sinauer Associates, Inc.
- Howe C.J. *et al.* (2001). Manuscript evolution. *Trends in Genetics*, vol. (17): 147-152.
- Lantin A.-C. (2003). *Phylogénie des manuscrits de Grégoire de Nazianze*. Final dissertation. UCL.
- Mooney L.R. *et al.* (2003). Stemmatic Analysis of Lydgate's Kings of England: a test case for the application of software developed for evolutionary biology to manuscript stemmatics. *Revue d'Histoire des Textes*, vol. (31): 202-240.
- Nei M. and Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.
- Macé C. *et al.* (2003). Le classement des manuscrits par la statistique et la phylogénétique: le cas de Grégoire de Nazianze et de Basile le Minime. *Revue d'Histoire des Textes*, vol. (31): 241-273.
- O'Hara R.J. and Robinson P.M.W. (1993). Computer-assisted methods of stemmatic analysis. *The Canterbury Tales Project Occasional Papers*, t. (I): 53-74.
- Page R.D.M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, vol. (12): 357-358.
- Swofford D.L. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.