Dynamic Homology and Circularity in Cladistic Analysis

Ariel Jonathan Roffé
CEFHIC-UNQ-CONICET; UBA; UNTREF
ORCiD: https://orcid.org/0000-0002-0051-2028
ariroffe@hotmail.com / arielroffe@filo.uba.ar

Abstract

In this article, I examine the issue of the alleged circularity in the determination of homologies within cladistic analysis. More specifically, I focus on the claims made by the proponents of the dynamic homology approach, regarding the distinction (sometimes made in the literature) between primary and secondary homology. This distinction is sometimes invoked to dissolve the circularity issue, by upholding that characters in a cladistic data matrix have to be only primarily homologous, and thus can be determined independently of phylogenetic hypotheses, by using the classical Owenian criteria (for morphological characters) or via multiple sequence alignment (for sequence data). However, since in the dynamic approach, sequence data can be analyzed without being pre-aligned, proponents have claimed that the distinction between primary and secondary homology has no place within cladistics. I will argue that this is not the case, since cladistic practice within the dynamic framework does presuppose primary homology statements at a higher level.

Keywords: Homology; Circularity; Cladistics; Dynamic Homology; Multiple Sequence Alignment; Direct Optimization

1. Introduction

A classical problem in the philosophy of systematics is the alleged circularity in the determination of homologies (see Jardine 1967; Patterson 1982; Wiley and Lieberman 2011, p. 117; Blanco 2012, among many others). It is usually acknowledged that not every trait is

useful or informative for recognizing phylogenetic relationships. Particularly, traits that are identified as "the same" across species because of mere functional similarity (i.e. analogies, for example the possession of wings in birds and insects) are not indicative of common ancestry, only traits that are homologous are. Thus, homologies have to be identified prior to the beginning of phylogenetic analysis itself (e.g. characters and states in a cladistic data matrix[1] have to be homologous). On the other hand, homologies are sometimes defined as those traits which are derived from a trait in a common ancestor (see for example Simpson 1961, p. 78; Mayr 1982, p. 45; Futuyma 2005, p. 49; Pearson 2010, pp. 483-484, among others), which implies that phylogenetic relationships (and thus phylogenetic analysis) have to be established before homology recognition can take place. Hence, phylogenetic analysis seems to require knowledge of its own output in order to take place. To put it differently, phylogenetic analysis rests on the assumption that common ancestry *explains* the presence of homologous traits. But, if homologies are *defined* as those traits that are derived from a common ancestor, then (by replacing the concept of homology with its definition in the first sentence) we get that common ancestry explains the presence of traits that are derived from a common ancestor, a clearly circular claim.[2]

There is a vast literature, both in the philosophical and biological literatures, on the way of resolving (or dissolving) this apparent issue. One proposal that has gained widespread acceptance is that of de Pinna (1991), which, if not the standard solution, has become at least an obligated reference in discussions on homology within systematics. According to de Pinna, one must distinguish between two senses of homology used in systematics. "Primary" homologies are the characters and states that are similar in more than a purely functional way. They are recognized using the classical Owenian criteria of topographic correspondence, composition, etc. (developed originally by Saint-Hilaire 1830, and Owen 1848, 1849 and synthesized in Remane 1952; see also Brady 1985; Caponi 2015). These are the data that go into the input data matrix, and which do not presuppose knowledge of the

[1] In this article, I focus on cladistic (i.e. maximum parsimony) approaches to phylogenetic reconstruction, leaving out other currently used methods such as maximum likelihood and Bayesian analysis. However, much of what I will say here will be applicable to those methods as well, since the dynamic homology approach has been applied to them as well (e.g. Wheeler 2006 for likelihood, Herman et al. 2014 for Bayesian inference). With the term 'cladistics' I am referring solely to the method of phylogenetic reconstruction, not to any views about classification (see Quinn 2017 for the various uses of the term).

[2] A similar point has been raised regarding the concept of adaptation, see e.g. Ginnobili (2018, pp. 26-31) and references therein.

phylogenetic relations among the taxa under study in order to be operationalized. On the other hand, "secondary" homologies are those traits (character states) that are derived from a common ancestor and which are, thus, part of the output of phylogenetic analyses. These analyses typically show that not every primary homology is in fact a secondary homology. That is, that some structurally similar shared traits are in fact homoplasious (convergent, etc.).

In analyses that use sequence (DNA or aminoacid) data as characters, the preliminary step of primary homology recognition (i.e. the establishment of correspondences among the nucleotide bases from different sequences) is carried out through the use of multiple sequence alignment (MSA hereafter). In MSA, a cost is assigned to different kinds of nucleotide substitutions and to indels, and based on that, gaps are inserted into the sequences in order to minimize the sum of the cost (see the next section for more details).

The two-step procedure described thus far, where phylogenetic analysis is divided into two successive stages (primary homology recognition and phylogenetic analysis proper), is the way in which the vast majority of the work within the cladistic paradigm has been carried out since its inception, and is carried out today. However, starting in the mid-1990s, a number of researchers have been working on an approach to phylogenetic analysis that (supposedly) obviates the need to perform the first step in that procedure. This general approach to phylogenetic analysis has received the name of the "dynamic homology" approach (contrasted with the "static", traditional, approach). Particularly in the context of sequence data, Ward Wheeler has devised a series of techniques (and implemented them in a computer program called POY, see Wheeler et al. 2015) that allow one to perform a cladistic analysis using sequence data without having them pre-aligned (Wheeler 1996, 1999, 2003a, b; the first of these techniques is described below in section 3). This has led Wheeler and his colleagues to claim that the distinction between primary and secondary homology makes no sense (Wheeler et al. 2006, p. 10; see also Vogt 2018, p. 219). More specifically, that cladistics has no need for what de Pinna had called primary homologies, and that the only concept of homology present within systematics is de Pinna's secondary homology.

In this article, I examine what the dynamic approach entails for the discussion about homology and circularity within systematics, and more specifically, whether Wheeler et al.'s claims about the distinction between primary and secondary homology are adequate. My

perspective here will be metatheoretical. That is, I will neither defend nor attack the dynamic homology approach (though I sympathize with it and will mention the advantages that its proponents believe it has). On the contrary, my goal will be to examine *if* adopting it entails that one should abandon de Pinna's distinction. I will uphold that the practice of systematics within a dynamic framework still implicitly presupposes a preliminary step of primary homology recognition, not between particular bases, but between the whole sequences. A secondary goal is simply to bring attention to these developments, since they have potentially important consequences for philosophical issues but have thus far not been examined at all by philosophers.

In the remainder of this article, I will proceed as follows. In section 2, I will describe the traditional approach to cladistic analysis in much greater detail, explaining why circularity can be thought to arise within this approach and how it is usually dealt with, both conceptually and operationally (specifically in DNA sequence data). In section 3, I will present the dynamic homology approach, and more specifically Wheeler's Direct Optimization technique (hereafter DO; previously referred to as Optimization Alignment), as well as the advantages its proponents claim it has. In section 4 I will analyze Wheeler et al.'s claims regarding the concept of primary homology and argue that one may still find a step of primary homology assessment implicit in dynamic practice. Section 5 considers two possible objections to this last point. Finally, I will draw some conclusions.

2. The Classical Two-Step Approach to Cladistic Analysis

In this section, I briefly describe the traditional, two-step approach to cladistic analysis, as well as the (by now classical) distinction between primary and secondary homology taken to be implicit in this methodology. This introduction will be short and to the point, readers looking for a more in depth understanding of cladistic analysis may see Kitching et al. (1998) and Wiley and Lieberman (2011).

2.1 Cladistic Analysis: The Basics

What cladistic analysis seeks to uncover is the pattern of diversification among a set of taxa. For instance, for the species S_1, S_2, S_3 and S_4, the following are two (out of 15) possible branching patterns (figure 1):

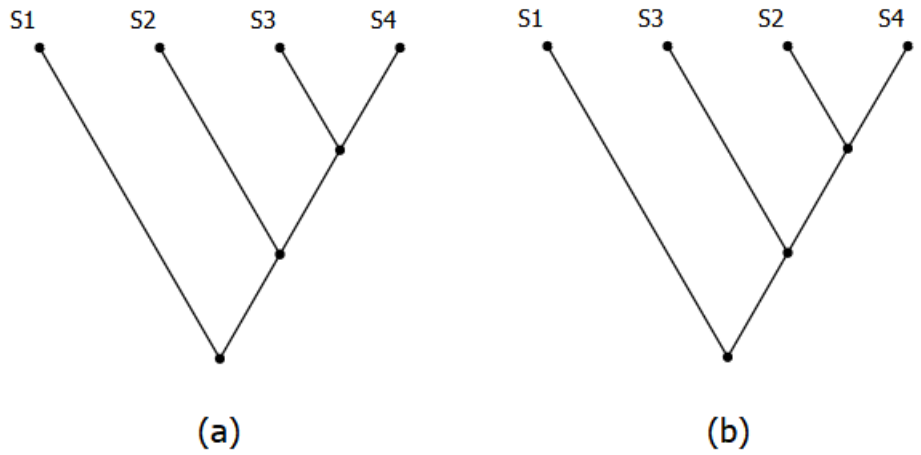


Fig. 1 Two possible branching patterns for four species. All trees in this article were drawn using https://www.phytreemaker.com/

Branching patterns are represented via rooted trees, called cladograms, where each taxon of interest is placed at a terminal node (hence, they are sometimes called terminal taxa). Internal nodes represent hypothetical ancestors.[3] Thus, according to the first cladogram, S_3 and S_4 share a common ancestor that is not an ancestor of S_2, while in the second cladogram that is the case for S_2, S_4 and S_3, respectively.

To pick a tree among the set of all possible trees, cladistic methodology looks at the distribution of traits (characters and character-states) among the terminal taxa and chooses the tree (or trees, see below) that best explain these "observed" data. In more precise terms, cladistic analysis must be given a character data matrix, that may look like this (table 1):

	C_1	**C_2**	**C_3**

[3] This corresponds to what Martin et al. 2010 call a node-based reading of phylogenetic trees; those authors show that there are other, equivalent, ways of interpreting trees. There has also been some discussion over whether cladograms and phylogenetic trees are equivalent, and what each represents (see e.g. Platnick 1977; Wiley 1979). I shall not go into these nuances, since they will not affect my points.

S_1	0	1	0
S_2	0	0	0
S_3	1	0	0
S_4	1	1	1

Table 1. Example data matrix with 4 taxa and 3 characters.

What this matrix indicates is that, for character C_1, S_1 and S_2 possess state 0 (say, white flowers), while S_3 and S_4 possess state 1 (say, purple flowers). To choose a tree, each character is mapped (or in technical jargon, optimized) into each tree, to see how many evolutionary changes must be postulated to explain its state-distribution. Optimizing a character on a tree means assigning the character's states to the internal nodes in such a way as to minimize the number of state-transformations. For instance, character C_1 from this matrix can be optimized in the two trees presented above in the following way (figure 2):

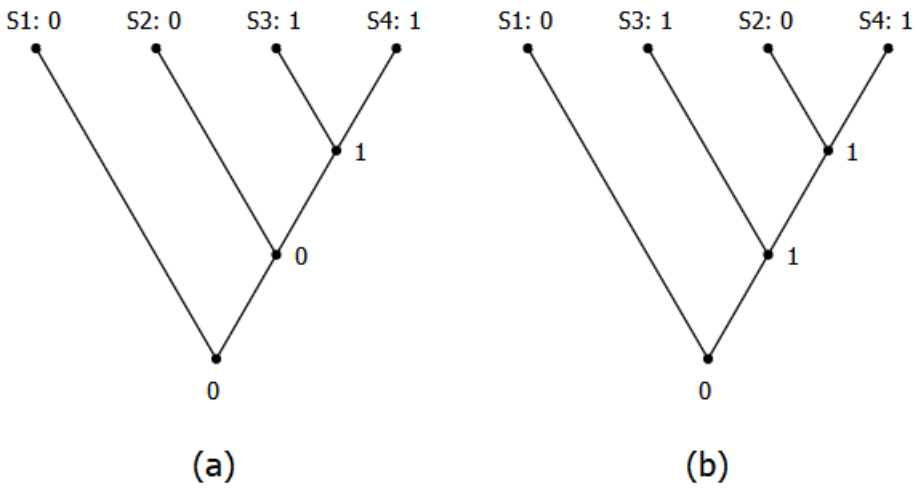


Fig. 2 Branching patterns from Fig 1. with character C_1 optimized on them. The first tree requires only one transformation (i.e. has a length *l* of 1), while the second requires at least two ($l = 2$).

If in the first tree (figure 2a) state 0 was put in the parent node of S_3 and S_4, then the character would not be optimized, since we would be (unnecessarily) postulating two changes instead of one in the transition from itself to its two descendants. The *length* of a cladogram is just the sum of the cost of the optimizations of each character, and the chosen tree (or trees) is

the one that minimizes length—i.e. that requires postulating the least amount of evolutionary change;[4] this is the reason why the method is sometimes called "maximum parsimony" (see Sober 1988).

There are many additional complications to the picture just presented (some of which will be discussed below). In the remainder of this subsection, I focus on one that will be important moving forward. Strictly speaking, it is not true that cladistic analysis always seeks the tree with the minimum *number* of changes. That is because, in some cases, some changes can be thought to be more costly than others. That is, up until now, I have been assuming that every transformation in every character counts the same (i.e. adds 1 to the cost of optimizing that character). But that is not necessarily the case. In some cases, some transformations can be taken to be more costly than others, either within a character or between characters. For instance, a major modification in morphology could be assigned a greater cost than a change in a minor feature. In DNA sequence data, indels can be assigned a greater cost than substitutions, and transversions can be given a greater cost than transitions (in each case, reflecting the lower or greater frequency with which they tend to occur in nature)

The transformation costs can be specified via a |C| × |C| cost matrix for each character C (where |C| is the number of states in C). For instance, the cost matrices for characters C_1 – C_3 from above could look like this (tables 2, 3, 4):

C_1	**0**	**1**
0	0	1
1	1	0

C_2	**0**	**1**
0	0	2
1	2	0

C_3	**0**	**1**
0	0	3
1	1	0

Tables 2, 3, 4. Example alternative cost matrices for C_1, C_2 and C_3.

Here, C_1 would be as before. (Every change of state between two adjacent nodes adds 1 to the length of the tree.) However, C_2 would be said to weight twice as much as C_1, since two transformations in C_1 would be equivalent to one in C_2. (i.e. The method would prefer a tree with 3 changes in C_1 than another with two changes in C_2.) In the case of C_3, note that *within*

[4] If more than one tree has the same minimum length, then cladists will not typically choose one among them arbitrarily. Instead, they will present some consensus tree between all the optimal trees. For instance, the *strict* consensus tree will only contain the groups that are present in all the optimal trees (see Kitching et al. 1998, chapter 7).

it some transformations are more costly than others (e.g. moving from state 0 to state 1 is more costly than the reverse). Note that only the *relative* costs matter (how much a transformation costs *against others*), not their absolute values; that is, the units do not matter. Alternative cost matrices may yield different minimum length trees.

2.2 Cladistics and Homology, and the Problem of Circularity

Now that I have given a basic idea of how cladistic analysis works, let us consider the relations between cladistics and homology, to see how the problem of circularity is thought to arise within cladistics.

As explained above, finding the optimal cladogram first requires optimizing all characters in it, to measure its length. A character optimization on a tree gives us a scheme of homologies for that character and that tree. For example, in figure 2a, state 1 of character C_1 is homologous between S_3 and S_4, since their most recent common ancestor possesses that state, while the same state is not homologous between them in figure 2b, but rather convergent. Therefore, finding an optimal tree also means finding an optimal scheme of homologies.[5] In that way, a homology scheme (or schemes, see footnote 5) is one of the outputs of cladistic analysis. Moreover, as the example just considered shows, there is no way of knowing if a trait is homologous or convergent between two species without knowing what the evolutionary tree looks like (i.e. without having performed a phylogenetic analysis).

We already have one of the components of the circularity charge mentioned in the introductory section. To get the other component, consider the following. Besides the cost matrices, the results of cladistic analysis are obviously also very sensitive to the choice of characters and the assignment of their states to the terminal taxa. There are many (perhaps infinite) ways of "dividing" organisms into sets of characters and states, most of them resulting in spurious phylogenetic groupings. Thus, care must be taken when selecting a set

[5] Again, there are many possible complications here, which do not always allow one to get a single optimal homology scheme. For instance, as mentioned above, cladistic analysis can yield more than one optimal tree, and those different trees can imply different homology schemes. Additionally, even a single tree can imply more than one possible scheme of homologies. For example, the tree in figure 2b can also be optimized by putting a 0 in all internal nodes, changing the homology scheme implied by it. I will not dig any deeper into these issues, since the general point made above will be enough for my purposes.

of traits, coding them into characters and states, and assigning states to the terminal taxa (Rieppel and Kearney 2002, 2007; Sereno 2007).

A constraint usually put in place for the choice of characters is that they must be homologous. This means two different things. On the one hand, it means that all states within a character must all be possible forms of *the same* trait type. This is important because, once optimized in a cladogram, they will form a transformation series—i.e. all current states will be shown as modifications from some past state within the character. This is why all the states within a character are sometimes called *transformational* homologies (Patterson 1982). On the other hand, a second requisite is that the traits of organisms which are assigned the same state have to be *taxic* homologies (i.e. all the characters marked with a 1 within a column must be homologous with each other, but not with the ones codified by something different than a 1).

In this sense, it would seem that homologies have to be identified in order to construct the data matrix, and thus, a homology scheme is one of the inputs of cladistic analysis. Therefore, since a homology scheme is a necessary input of cladistic analysis, but (as said above) it can only be known as the result of those analyses, circularity seems to arise. Moreover, the circularity involved would not only be conceptual but also, and perhaps more problematically for practicing biologists, operational, since phylogenetic analyses would be impossible to carry out (carrying them out would require knowledge of their own output).

One popular solution to this problem is to claim that the homologies that have to be present in the data matrix are not the same than those that result from cladistic analysis. Thus, a distinction is made between "primary" and "secondary" homologies.[6] Characters and states in the data matrix are only primarily homologous, meaning something like similar in more than a purely functional sense. Cladistic analysis outputs a scheme of secondary homologies,

[6] This terminology is by de Pinna (1991), the article that has become an obligated reference on the subject. It can also be found (in some cases before de Pinna's publication, as he himself acknowledges) as homology vs homogeny (Lankester 1870); hypotheses of homology vs. homology (Patterson 1988); topographical correspondence vs homology (Rieppel 1988); etc. What I aim to show in this article is that dynamic analyses presuppose a data matrix that is built using the same (topological, compositional, etc.) criteria than those used in classical two-step analyses. Whether one calls the characters resulting from the application of these criteria homologies, primary homologies, hypotheses of homology or something that does not contain the term 'homology' at all (e.g. topographical correspondences) is irrelevant to my point. For simplicity's sake, I will continue using de Pinna's terminology.

that is, of shared traits inherited from a common ancestor.[7] Real analyses always involve some degree of homoplasy, that is, cases of traits that are primarily homologous but not secondarily homologous (hence, cladistic homoplasy and analogy are different concepts, the latter not necessarily being primarily homologous). Operationally, circularity is avoided because (supposedly, more on this later on) the recognition of primary homologies can be carried out independently of phylogenetic analysis itself (Roffé et al. 2018). Let us dig deeper into this last point, to see why this would be case.

2.3 Primary Homology Recognition in Sequence Data

As said above, for morphological traits, the criteria used to recognize primarily homologous traits are the classical Owenian ones. Sequences (DNA or aminoacid, for simplicity I stick to DNA from hereafter), which are the focus of this article, present a different kind of difficulty. Since DNA sequences are composed of the same four, massively repeated, building blocks (the nucleotide bases A, C, T and G), an A in one sequence could in principle correspond to many A's in a different sequence (since all A's are qualitatively identical). For this reason, an MSA is carried out to infer base-to-base correspondences, which constitute a set of primary homology claims (Giribet and Wheeler 1999, p. 132). This works as follows.

Suppose that S_1-S_4 have the sequences CTATC, CGTAC, CGTTC and CAC, respectively, for a given gene. Since these sequences have different lengths, gaps gave to be inserted somewhere to establish which nucleotide base corresponds to which in the other sequences. A gap will represent the occurrence of an insertion or deletion event in some of the sequences (or indel for short; there is no *a priori* way to tell which of the two is the case). The following are three different ways of doing that (tables 4, 5, 6):

[7] To clarify, in my view, secondary homologies are traits derived from a common ancestor, and cladistic analysis gives one way (but not the only way) of *operationalizing* this concept. Others might wish to directly define secondary homologies as the output of the method. Although interesting, this discussion has no impact on the conclusions drawn in this paper.

S_1	C	-	T	A	T	C
S_2	C	G	T	A	-	C
S_3	C	G	T	-	T	C
S_4	C	-	A	-	-	C

S_1	C	T	A	T	C
S_2	C	G	T	A	C
S_3	C	G	T	T	C
S_4	C	-	A	-	C

S_1	C	-	T	A	T	C
S_2	C	G	T	A	-	C
S_3	C	G	T	-	T	C
S_4	C	-	-	A	-	C

Tables 4, 5, 6. Three possible alignments for the sequences CTATC, CGTAC, CGTTC and CAC. Each alignment implies a primary homology scheme. For example, the A in S_4's sequence is homologous to the A in S_1's sequence in tables 5 and 6, but not in table 4.

The first of these alignments contains three indels and one substitution, the second contains two indels and three substitutions and the third contains four indels and no substitutions. To choose between them, a cost must be given to each kind of substitution as well as to indels. (These are alignment costs, not phylogenetic transformation costs, so they may or may not be equal to the ones used later on during phylogenetic analysis. See section 3 for more on this point). For example, if indels are assigned a cost of 3 and substitutions (of all kinds) a cost of 1 (again, only relative costs matter, see above), then the second alignment will be preferable, with a total cost of 9 (vs. 10 and 12); however, for c(indel) = 3 and c(substitution) = 2, the first alignment will be preferable among these three, with a total cost of 11 (vs. 12 for the other two). Once the alignment costs have been chosen and an optimal alignment is found, this alignment can be used directly as (part of) a cladistic data matrix. That is, the columns of the resulting alignment will be characters in the phylogenetic analysis that follows.

Notice that if c(indel) = 2 and c(substitution) = 1, the first two alignments will be equally costly at 7. This can be problematic in practice because alignment programs will usually only return one alignment. However, two alignments that are equally costly *a priori* can generate optimal trees that differ in length in the following phylogenetic analysis. For instance, if the first alignment is chosen and the cost scheme is maintained identical (2 for indels, 1 for substitutions)—as it should be, see below in section 3—, then each of the 15 possible trees is optimal, having a length of 11 (each requires 5 indels and one substitution). However, the second alignment has only 5 optimal trees, which are the possible rootings of this unrooted tree (figure 3a):

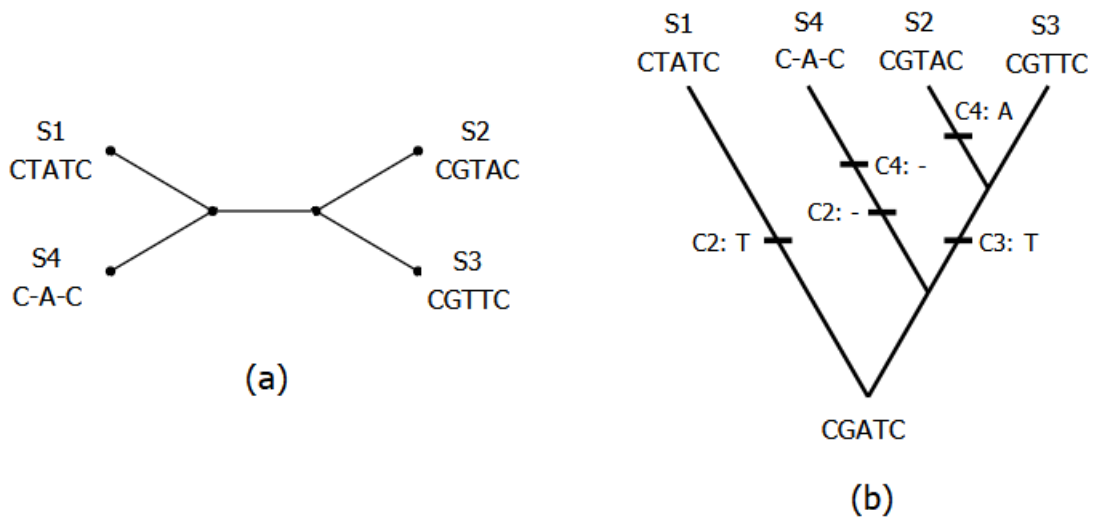


Fig 3. (a) Unrooted optimal tree for the alignment in table 5; **(b)** One possible rooting (on the branch leading to S_1) and optimization of the tree in (a).

each requiring only 3 substitutions and 2 indels, with a length of 7. Thus, the second alignment is preferable to the first, because it leads us to consider evolutionary scenarios that require postulating a less costly sum of evolutionary events. But this can only be known *a posteriori*, after a phylogenetic analysis has been made. (All of this is related to some of the critiques from the dynamic approach, see section 3.) As we shall see below, there are still other possible alignments that generate optimal trees of the same length, and that imply other possible secondary homology schemes.

Summing up, the general picture up to this point is the following. Cladistic analysis always proceeds in two steps. In the first step, one identifies primarily homologous traits (i.e. correspondences between morphological structures or between base pairs) based on criteria that are independent of phylogenetic considerations. In the case of sequence data, this is achieved through an MSA. In the second step, which we could call phylogenetic analysis proper, the optimal tree(s) for the data matrix elaborated in the first step is found, with its corresponding secondary homology scheme(s). In this way, since the data matrix is built independently from phylogenetic analysis proper, circularity is (at least operationally) avoided. In the next section, I introduce the dynamic homology approach, as well as the critiques that its proponents make to this general picture.

3. The Dynamic Homology Approach and Its Advantages

The dynamic homology approach is a heterogeneous set of methods and techniques for analyzing data of different kinds (from sequences, Wheeler et al. 2006; to morphology, Ramírez 2007, Agolin and D'Haese 2009, Vogt 2018; to behavior, Robillard et al. 2006 Japyassú and Machado 2010; in this article I focus mainly on sequences). The unifying idea within this collection is to not have primary homology statements coded and fixed in a character data matrix, but rather to build those characters in the course of phylogenetic analysis.

To get an idea of what this means, consider one criticism that the dynamic approach proponents make to the static approach: that primary homology hypotheses are not revised once they are placed in the data matrix (e.g. Wheeler 2001b, p. 304; Wheeler et al. 2006, p. 30). This may seem strange given what was said previously, that phylogenetic analysis typically reveals that many traits coded as the same state turn out to be homoplasious. Indeed, taxic homologies are revised in the standard procedure. The point is rather that *transformational* homologies are not. Once a set of states is placed within a column, they will form a transformation series in the resulting optimized tree. In other words, *characters* are not revised during tree search, when, sometimes, a different character coding could lead to different, more globally optimal solutions (see the example in the last subsection above).[8]

What Wheeler realized was that sequences need not be pre-aligned but could rather be aligned on the candidate trees during tree search. To understand this, consider how MSA usually works in practice. For two sequences, there is an algorithm by Needleman and Wunsch (1970) that provides an exact solution (i.e. is guaranteed to arrive at a lowest cost alignment), and that works reasonably fast. However, once more sequences are included, the problem becomes computationally intractable. This is why alignment programs use heuristic

[8] In a static setting, characters can be revised *after* phylogenetic analysis has taken place, by recoding the data matrix that was used as its input and analyzing it again. This is a common characteristic of all science: one can revise the data after seeing how well they fare against the theoretical analysis of it (confirmation holism taught us this long ago). But what dynamic analyses do is different, they are effectively changing the theoretical analysis itself, and how it treats the data.

methods, which are fast and provide good solutions, but which are not guaranteed to be the globally optimal ones (i.e. there may be other alignments of lower cost that the method misses).

These heuristic methods work by decomposing a hard problem (the MSA) into many simpler problems (aligning two sequences), by performing a series of pairwise alignments guided by a binary tree (Wheeler 2001b, p. S4). This binary tree seems, by all accounts, to be a phylogenetic hypothesis. Thus, the idea that base-to-base correspondences (primary homologies in DNA sequences) are determined independently of phylogenetic hypotheses seems false, threatening to bring back the issue of circularity. Moreover, the guide tree is usually obtained using distance methods, which are considered to be less adequate than parsimony or probabilistic methods for the purpose of phylogenetic reconstruction. So, an additional criticism to the static approach would be that phylogenetic hypotheses assumed in primary homology assessment are obtained from inadequate methodologies, with possibly cascading effects later on.

Instead, Wheeler's (1996) DO method uses each candidate tree in a cladogram search as a guide tree. That is, no *a priori* alignment is used to evaluate tree length, but rather a (possibly different) alignment is generated for each tree, by using a similar pairwise alignment algorithm. Therefore, the DO method chooses the most globally optimal tree *plus* alignment scheme. To see how this works, consider the following example (a more complete description of the DO method can be found in Wheeler 2002). Take the four sequences from the end of section 2, the cost regime c(indel) = 2, c(substitution) = 1, and the following tree (figure 4a):

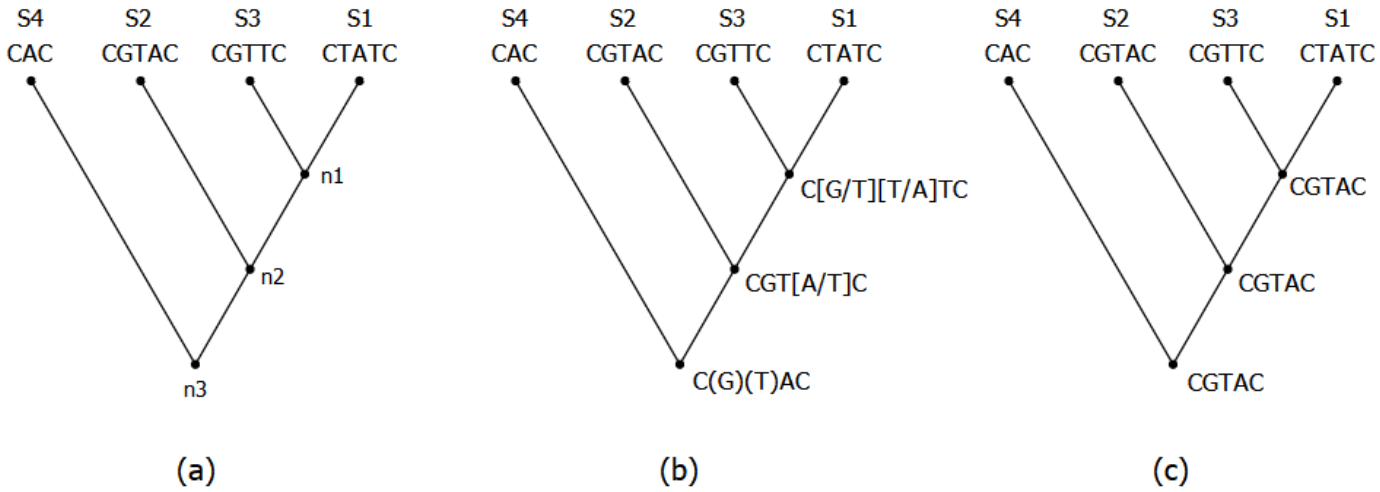


Fig. 4 Example tree (a), and four sequences optimized using the DO method in (b) and (c), see the description below.

The DO method will first perform a down-pass (figure 4b) and then an up-pass (figure 4c) to find the sequences of the ancestors, and thus the length of the tree. For the down-pass, first, the sequences of S_1 and S_3 will be aligned, and its result will be the (preliminary) state of node n_1. CTATC and CGTTC can be aligned into C[G/T][T/A]TC; the bracketed positions mean that any of the bases could be present in the ancestor with the same cost (e.g. for the second position, if a G is present, then a substitution will have occurred in S_1 with cost 1, and if and a T is present, the substitution will have occurred in S_3 with cost 1). In other words, the sequence of the ancestor may have been either CGTTC, CGATC, CTTTC or CTATC. The down-pass proceeds by aligning n_1's sequence, C[G/T][T/A]TC, with S_2's, CGTAC. The optimal way to do so is to consider G and T to be the case in n_1's second and third position, and postulating only one substitution in the fourth, yielding the sequence CGT[A/T]C. Finally, this sequence is aligned with S_4's CAC to get the sequence of the root node n_3. Here the sequences have different lengths, so gaps will necessarily have to be inserted somewhere. The optimal way to do so is to pair the initial and final C's, and to consider n_2's fourth position be an A, thus requiring only two indels and no substitution. In the resulting sequence C(G)(T)AC, a position of format (G) means that the ancestor may or may not have contained that base. That is, the ancestor could either have been CGTAC, CGAC, CTAC or CAC.

For the up-pass, each ambiguous node is considered from the bottom up, but now with regards to its three surrounding nodes (except for the root, which can always be resolved

in any way). For instance, we had already established that the second position in n_1 could either be a G or a T. If it is a T, then two substitutions will have occurred (one below it from CGT[A/T]C and one above it to CGTTC at the left), while putting a T involves adding only one substitution (to CTATC at the right). Thus, this ambiguity resolves to a G. One possible resolution of all nodes is shown in figure 4c (there are others, however).

All optimal resolutions in the up-pass require postulating the same number of evolutionary events, two indels and three substitutions in this case. The length of the tree can thus be calculated in the down-pass. Again, in this case, this will equal 7 with the chosen cost regime. Notice that this is the same minimum length that we had found using the classical approach in section 2. However, the topology in figure 4 was not recovered as optimal in that section. To see why, trace the evolution of the individual base positions on the tree in figure 4c (in the way described in Wheeler 2003c) to get the "implied alignment" from this tree (table 7):

S_1	C	T	A	T	C
S_2	C	G	T	A	C
S_3	C	G	T	T	C
S_4	C	-	-	A	C

Table 7. Implied alignment from the tree in figure 5c.

As the reader may see, this is almost identical to the alignment in table 5, with the A in S_4 moved to the right. However, this slight change makes that character group S_4 with S_2 instead of with S_1 in C_3, thus changing the optimal cladogram.

In this sense, we can see what the dynamic approach proponents mean when they criticize how characters are not revised in standard methodology. In this case, sticking with only one *a priori* alignment will inevitably make us miss some optimal trees (and possible homology schemes). In real, more complex, cases, where the guide tree used in the MSA tends to be somewhat inadequate, DO pretty much always returns lower cost trees than the standard methodology.

Finally, we can mention one more practical and one more conceptual advantage of the dynamic approach against the static one. The first is that, in the vast majority of actual

cases, the cost schemes used for alignment and for tree search do not coincide (furthermore, in many cases, gaps are treated as missing data instead of a fifth state during tree search). This introduces some conceptual incoherence, since the same biological phenomena (substitutions and indels) are weighted differently in different parts of the analysis. The practical advantage of the dynamic approach is that it makes this impossible.

The conceptual advantage concerns the way to conceptualize gaps. Gaps are not real or "observed" states, in the sense that sequences never actually contain a "-" base. Rather, gaps signify evolutionary events of either insertions or deletions relative to the sequence of an ancestor (Giribet and Wheeler 1999). Thus, it is reasonable to think that evolutionary events should only be postulated during a phylogenetic analysis and not be assumed prior to it. To assume that certain evolutionary transformations have taken place before phylogenetic analysis is carried out can lead to bias in the selection of a tree, as shown by the above examples (using either the alignments in table 5 or 7 will make us miss some optimal trees, while using the alignment in table 4 will return all trees, some of which are actually not optimal globally).

To sum up, one of the methodologies within the dynamic approach, DO, works by using each candidate tree during tree search as a guide tree for an alignment, instead of taking a fixed pre-aligned set of characters. As shown above, this has many advantages, which include recovering more and/or lower cost trees and homology schemes, avoiding conceptual incoherences, treating gaps more adequately, etc.[9] In the next section, I consider the alleged consequences that this approach has for the distinction between primary and secondary homology.

4. Primary Homology in the Dynamic Approach

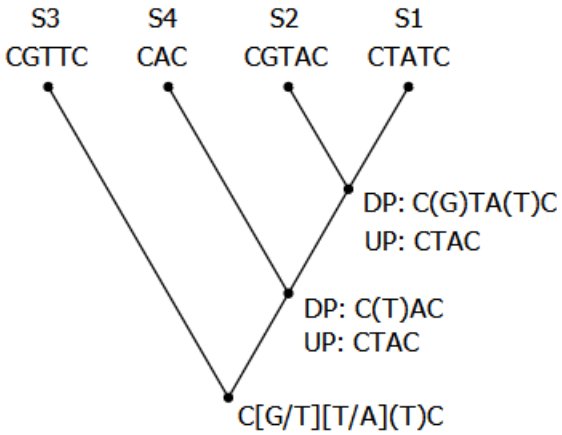
Having explained the static and dynamic homology approaches in the last two sections, what will interest me in the rest of the article are the consequences that the dynamic proponents draw for the distinction between primary and secondary homology. According to Wheeler

[9] The main (practical) disadvantage of the DO method is that it is computationally much more intensive than the traditional methodology.

and his colleagues, the distinction has to be discarded. The fact that sequences do not need to be pre-aligned would show that there is no need for the establishment of primary or putative homologies. The only concept of homology necessary for cladistics is de Pinna's secondary homology. In their words:

> (…) [C]ladograms imply statements of homology. Alternative cladograms might have alternative optimal homology statements and content (…) Features are homologous when their origins can be traced to a unique transformation on the branch of a cladogram leading to their most common recent ancestor. There can be no notion of homology without reference to a cladogram (…) This definition of homology makes no reference to 'primary' or 'secondary' homology (de Pinna 1991). In fact, the perspective here rejects this distinction entirely. (Wheeler et al. 2006, p. 10)

In support of this view, one could claim that although base-to-base correspondences are established during tree search, the resulting matrix (the implied alignment) is a secondary homology scheme among bases, not a primary one. To see this clearly, consider the same four sequences from above, a cost scheme of 4:3 for indels vs substitutions and the following optimized tree (figure 5a):



(a)

S_1	C	-	T	A	-	T	C
S_2	C	G	T	A	-	-	C
S_3	C	-	G	T	T	-	C
S_4	C	-	-	A	-	-	C

(b)

Fig 5. (a) Tree optimized with the four sequences from section 3 (DP: down-pass optimizations, UP: up-pass resolutions). **(b)** Implied alignment from (a)

The implied alignment for this optimized tree is shown in figure 5b. As the reader may see, some columns (e.g. five and six) are absurd from an MSA standpoint, since putting the two T's together in the same column would reduce the alignment cost by an indel. However, they are separated since these two T's came up through two different insertions according to the evolutionary hypothesis depicted in the tree (they do not form a transformation series). Hence, this alignment depicts a secondary homology scheme.

On this point, I believe Wheeler et al. are right. Phylogenetic analyses done within the dynamic frame have no need for a primary homology scheme *among particular nucleotide bases*. However, this does not imply that there is no need for primary homologies *at all* within this approach. In what follows, I will argue that primary homology statements can still be found, but at a more general level, i.e. the level of entire sequences

To see this, remember that the distinction between primary and secondary homology allowed us to see how phylogenetic analyses avoid operational circularity, by showing that the data matrix (the input of the phylogenetic analysis proper) can be constructed independently of the phylogeny that is the output of the analysis. To avoid this circularity problem, this point must still hold: the input data still have to be collected independently of the result of the analysis itself. The question is now what the input data are in a dynamic setting, and how they are gathered. The answer lies in recognizing that the data matrix in a dynamic setting contains one character per *locus*, not per nucleotide base.[10] For example, in the case analyzed in the last section, the data matrix looks as following (table 8):

	C_1
S_1	CTATC
S_2	CGTAC
S_3	CGTTC
S_4	CAC

Table 8. Data matrix presupposed in the example from section 3.

[10] I am not claiming that individual sites are not characters, in the wider sense of characteristics that the study taxa possess, and that can therefore be compared and homologized (see above). I only state that, in a DO setting, the input matrix does not contain site-characters but rather *loci*-characters.

As explained in section 2, there are some constraints on what can be placed in the data matrix. In the same way that the possession of wings would constitute a poor character in an analysis that includes insects and birds, taking random parts of the organisms' DNA and placing them in the above matrix would yield a completely worthless result. The entire sequences have to be *the same* in some sense, i.e. homologous. That is, they have to belong to the same *locus*, and thus form a transformation series which will, itself, not be revised in the course of the analysis. The question now is which kind of homology statement this is. Is it a primary or a secondary homology statement?

A way to answer this question is to look at the operationalization criteria used to recognize them—i.e. how we recognize that two sequences belong to the same *locus* or gene in two different species of organisms. If these correspondences are inferred from a previous phylogenetic analysis of the species in question, then the statement could be claimed to be a secondary homology statement. Otherwise, if classical similarity criteria are used, then we should conclude that this is a primary homology statement.

A case where the answer to this question is most clear is when a new species is found, and researchers want to perform an analysis to place it within a phylogeny. In that case, clearly, there is no previous phylogeny available, so the determination of the sequence for the desired *locus* cannot possibly rely on previous phylogenetic hypotheses. Let us dig deeper into how exactly the determination of the sequence would take place.

One way to find the homologous sequence in a new species would be via a PCR. In a PCR, an entire DNA fragment + small regions of DNA (called primers) that surround the desired sub-sequence are given, and (through a complicated process which I shall not go into here) many copies of the region between them are generated, which can later be sent for sequencing. The strategy consists in taking primers from the desired *locus* of some already known species, and then use them in a PCR with the new species' DNA. In other words, if S_1 is a known species and S_2 a new one, we find the desired sequence in S_2 by sequencing the part of its DNA that is surrounded by small fragments similar to those that surround S_1's gene. This is clearly a topographic criterion (being surrounded by similar things), which suggests that the homology in question is primary.

An actual example (among countless others) where this kind of procedure was used can be seen in Ceccarelli et al. (2019). This study established a phylogeny of a group of spiders, in order to study the evolution of various adaptations to living in grasslands. For this, it contained four genetic regions (two nuclear and two mitochondrial) in its input data matrix. Of the 119 spider taxa considered, 38 were new to it, and part of the DNA data was missing for five other taxa (which had been previously used in other phylogenetic analyses). To determine all this missing DNA data, the procedure described in the last paragraph was used. The supplementary files (Ceccarelli et al. 2018) contain the details of the PCR procedure employed, as well as a table (Table S2 in File 1) detailing the primers that were used for the four regions, and where they were taken from (e.g. for the 16S B gene, the primer CCGGTTTGAACTCAGATC was used, which was taken from Simon et al. 1994).

The conclusion that these are primary homology statements is even more obvious at a larger scale. A previous question to carrying the above procedure could be which species to compare the target organism with (i.e. to take primers from). For instance, once a new species is found, it will first be identified as a spider based on other similarity criteria. For example, it will possess some morphological traits that spiders typically possess (eight legs, spinnerets, etc.). Moreover, the relevant DNA piece to provide to the PCR will be located in a topographically similar region of the organism's cell (the nucleus or the mitochondria) and will be compositionally made of the same chemical elements, etc. These are all classical similarity criteria that have to be employed (and are usually implicit in practice) in the determination of the desired sequence, in order to build the data matrix from table 8.

To clarify, the cases where some of the sequences are unknown or missing provide the clearest and most transparent reasons for arguing that the sequences contained in a dynamic data matrix are primarily homologous. But this would also hold if all the DNA data were previously available. For instance, if one were building a phylogeny using the 16S gene, and the sequences for this gene were available in GenBank for all the study taxa, the very fact that they were *labeled as the 16S* in GenBank would already imply a primary homology statement. The researchers who originally uploaded the sequence in question probably did something like the procedure I described above. Surely, *we* (as the people making the phylogenetic analysis) would not be making these primary homology judgements because someone else would have made them for us. This is just standard division of labor.

The conclusion of this section is that, even though it is true that the dynamic homology approach does not require primary homology statements between particular nucleotide bases as input, phylogenetic practice done in its fashion does (implicitly) presuppose some primary homology statements at a higher level. Although these conclusions were drawn from the analysis of the DO approach, as stated in footnote 1, they also extend to dynamic proposals that use other optimality criteria, such as likelihood and Bayesian inference. This can be seen because the input data matrices used in those analysis are almost identical to the ones used in parsimony (i.e. aligned nucleotides in a static framework, entire *loci* in a dynamic one). Thus, primary homology statements are equally presupposed in those approaches in the construction of the data matrices.

These conclusions also hold for dynamic analyses that use morphological characters, but the reasons for that may be different, and may vary depending on the details of the proposal. For instance, in Ramírez's (2007) approach, dynamic considerations enter when the primary homology determination criteria (topography, composition, ontogeny, intermediate forms, etc.) give us conflicting matchings. For example, there is an ongoing debate about whether the digits in the avian three-digit limb correspond to our I-II-III or II-III-IV digits. Topographical and paleontological intermediate forms suggest a I-II-III match, but embryology suggests II-III-IV (see Young & Wagner 2011, and references therein). What Ramírez proposes is to build two (or more) data matrices with the different possible alternative codings, which imply different transformation series, and then analyze both in a classical way, choosing the one that requires less (weighted) transformations. In the case of the digit homology controversy, this was done by Xu et al. (2009). In this sort of approach, it is even easier than with DNA to see that primary homology statements cannot be dispensed with, because both matrices are built using the classical homology criteria. There is nothing like a "morphology alignment" happening during tree search, character identification and phylogenetic analysis happen at two separate steps, as in the static approach. Other morphological approaches, however, operate differently. For instance, Vogt's (2018) proposal (which is still programmatic and not fully operational yet) does not even use two-dimensional data matrices but rather semantic ontologies, and some form of similarity

judgments do happen during tree search, so the argumentation would need to be (but could be) adjusted.[11]

One last interesting point is that one could say that the static approach also presupposes the establishment of sequence-homologies, since homologous sequences must be given as input to the *a priori* alignment tools. In that sense, perhaps classical analysis is not only comprised of two steps but three (or more if one continues to go up levels), and the dynamic framework does involve one less step. This does not imply, however, that it can do without the primary/secondary homology distinction entirely, which is what I am arguing for here.

5. Some possible objections

In this section, I consider two possible objections to my arguments from the last section. The first would come from trying to restate Wheeler et al.'s position one level up. I've shown above that the non-necessity of identifying base-to-base primary homologies in a dynamic setting still presupposes primary homology statements at the level of entire *loci*. However, one could hold that (in principle at least, if not in practice) the correspondences between *loci* themselves could be dynamically established —i.e. *loci* be aligned, by providing the analysis longer stretches of DNA that contain many genes. Thus, one could (in principle at least) avoid making those primary homology judgements.

My reply would be that this misses the general point I'm making. The general point is that if an analysis aligns *something* (at any level), that is, decomposes a whole into parts and says which parts correspond to which other parts in other entities, then the wholes being aligned have to be "the same" (i.e. homologous). If one is dynamically aligning nucleotide bases, then the entire genes have to be homologous. If one parts from longer sequences (that contain many genes) and dynamically decomposes them into genes, then those longer sequences have to be homologous. Otherwise, throwing random (longer) stretches of DNA into the method will result in a worthless result. Even if one could place entire genomes in

[11] The conceptual and mathematical apparatus used by Vogt is rather complex, and a fuller introduction and examination of it fall outside the scope of this article.

the input matrix, those genomes would have to be primarily homologous —their operationalization would require sequencing entities which are topographically located at the same place within an organism's cell, are chemically composed of the same elements (e.g. DNA), etc. These would still be (implicit) primary homology statements.

The second possible objection could come from critics of the distinction between primary and secondary homology who appeal to Hennig's (1966) idea of reciprocal illumination. The idea is that there is a give-and-take between theory and observation, that advancements in one area (e.g. character analysis and coding) can shed light into another related area (e.g. phylogenetic analysis), which in turn, illuminates the first, creating a virtuous circle. Thus, the story goes, no sharp distinction can be drawn between primary and secondary homology concepts, and the circularity problem mentioned in the introduction would not really be a problem but is merely part of the normal dynamics of science.

These more or less vague ideas were developed in the context of the dynamic homology approach by Grant and Kluge (2009). They had previously asserted that:

> The *explananda*, **e**, [of cladistic analysis] are the character-states (…) of terminal taxa, which are explained by postulating a particular hypothesis of phylogenetic relationships (i.e., a hypothesis of cladistic and patristic relationships), **h**, in light of the background knowledge of descent, with modification, **b**. Together, **b** and **h** constitute the *explanans*. (Kluge and Grant 2006, p. 284, emphasis in original)

Farris (2008, p. 829) objected to this that part of the hypothesis repeats the *explanandum*, because individuating something as a character involves postulating it as a transformation series, something that is phylogenetically charged. Grant and Kluge identify this as the tautology objection introduced in sections 1 and 2 of this article:

> Thus, as portrayed by Farris (2008), by treating character-states as the *explananda*, *e*, and defining character-states as resulting from transformation events, which are themselves inferred through phylogenetic analysis and are therefore part of *h*, Kluge and Grant (2006) violated Popper's rule and thereby committed a fatal logical error.

> However, on closer inspection Farris' accusation is nothing more than the old claim that evolutionary inference is circular. This charge was famously aimed at Hennig by Sokal and Sneath (…). (Grant and Kluge, 2009, p. 359).

Their response to this objection is to hold that, in a dynamic setting, transformation series are not merely assumed, but rather they are tested in conjunction with the trees. Thus, tree assessment (the *explanans*) and character assessment (the *explanandum*) are evaluated together during tree search, one shedding light on the other, without any vicious circularity appearing.

What Grant and Kluge fail to see is that, in the dynamic approach, the *explanandum* (of phylogenetic analysis proper) changes. They are right in claiming that what is explained is the distribution of character states present in the data matrix. But, as said above, the data matrix contains entire sequences as states, not particular nucleotide bases. That is, in a dynamic setting, what one knows beforehand and tries to explain are the "observed" sequences, and what one postulates to account for them (the *explanans*) is a tree and an alignment. The hypothesis that the whole sequences form a transformation series is not tested or revised during a dynamic search, just as much as transformational homology statements between particular nucleotides are not revised in standard methodology.[12]

In any case, I should note that the general discussion (of whether, in general, the distinction between primary and secondary homology adequately captures phylogenetic theory and practice) is beside the point here. What I aimed to refute is the idea that if one adopts a dynamic approach then one is forced to abandon the distinction between primary and secondary homology (as Wheeler et al. claim), for reasons following from this adoption. If a researcher has independent reasons for rejecting this distinction, then my points do not concern them. What I have shown is that one can consistently adopt a dynamic framework

[12] All of this is not to deny that reciprocal illumination is a real phenomenon. As Hennig himself recognized (Hennig 1966, p. 21), it is not a phenomenon exclusive to phylogenetics. In all areas of science theories are tested through certain evidence, and at the same time the evidence can be modified to fit the theory. For example, if a theoretical prediction fails, one way to proceed is by revising the observations (either the initial conditions or the "observational" consequences), the measuring instruments, the auxiliary hypotheses, etc. In the philosophical literature this has long been known as confirmation holism. However, the occurrence of this phenomenon does not deny that the concepts in the *explanandum* of a theory have to be determinable independently of the theory for which they are *explananda*, on pain of circularity (as the structuralist school has long argued, and shown to be the case with multiple reconstructions of theories, see Balzer et al. 1987).

and at the same time maintain that primary homology assessment, based on classical criteria, plays a role in phylogenetic practice (at least to the extent that one can hold this to be the case in the traditional methodology).

6. Conclusions

In this article, I have evaluated the claim made by the proponents of the dynamic homology approach, concerning the adequacy of the distinction between primary and secondary homology. To review, this distinction is important because it enables one very popular way of responding to the classical circularity objection against phylogenetic analysis. The circularity objection claims that, since the data matrix must only contain homologies, on the one hand, and since a homology scheme is inferred from a phylogeny, on the other, then the input of cladistic analysis requires knowledge of its own output. One typical response to this objection is the aforementioned distinction—to hold that homologies that form the input of cladistic analysis are recognized using other, independent, criteria, and thus are merely primary or putative homology claims, which are later tested in the analysis itself.

Against this claim, the dynamic homology approach proponents have claimed that transformational homology statements are not actually tested in the standard methodology and that this can lead to bias in the selection of the trees and homology schemes. Instead, they propose various methodologies (one of which I presented in section 3) to infer phylogenetic relations and homology schemes which do not presuppose that sequences are pre-aligned. From this fact, they have gone on to claim that the distinction between primary and secondary homology makes no sense (at least within a dynamic setting), since everything is determined at the same time and using the same global optimality criterion.

I have argued that this last conclusion is not correct, since a preliminary step of primary homology assessment, using independent criteria, is still implicit in dynamic phylogenetic practice. This assessment is made at the level of entire sequences, not of particular nucleotide bases. As stated in the introduction, I wish to stress once more that this analysis does not entail any change to the practice of phylogenetic analysis, it only concerns the metatheoretical way in which we think about it and understand it. I would also reiterate

that the points made in this article do not constitute an argument against adopting a dynamic framework (which, again, I sympathize with). My only goal is to understand it better, so that the conceptual consequences we draw from it are more adequate.

Acknowledgements

I thank Martín Ramírez and three anonymous reviewers for commenting on an earlier version of the manuscript. This work has been funded by the research projects PUNQ 1401/15 and SAI 827-223/19 (Universidad Nacional de Quilmes, Argentina), UNTREF 32/15 255 (Universidad Nacional Tres de Febrero, Argentina) and UBACyT 20020170200106BA (Universidad de Buenos Aires, Argentina).

References

Agolin M, D'Haese CA (2009) An application of dynamic homology to morphological characters: direct optimization of setae sequences and phylogeny of the family Odontellidae (Poduromorpha, Collembola). Cladistics 25:353–385

Balzer W, Moulines CU, Sneed JD (1987) An architectonic for science: the structuralist program. Reidel, Dordrecht, Lancaster

Blanco D (2012) Primera aproximación estructuralista a la Teoría del Origen en Común. Ágora 31:171-194.

Brady R (1985) On the Independence of Systematics. Cladistics 1:113–126.

Caponi G (2015) El impacto de la filosofía anatómica de Étienne Geoffoy Saint-Hilaire em el desarrollo de la historia natural. Gavagai - Rev Interdiscip Humanidades 2:9–31.

Ceccarelli FS, Koch NM, Soto EM, et al (2018) Data from: The grass was greener: repeated evolution of specialized morphologies and habitat shifts in ghost spiders following grassland expansion in South America. doi: 10.5061/dryad.20257

Ceccarelli FS, Koch NM, Soto EM, et al (2019) The Grass was Greener: Repeated Evolution of Specialized Morphologies and Habitat Shifts in Ghost Spiders Following Grassland Expansion in South America. Syst Biol 68:63–77

de Pinna MCC (1991) Concepts and Tests of Homology in the Cladistic Paradigm. Cladistics 7:367–394

Farris JS (2008) Parsimony and explanatory power. Cladistics 24:825–847

Futuyma DJ (2005) Evolution. Sinauer Associates Inc., Sunderland, Massachusetts

Ginnobili S (2018) La Teoría de la Selección Natural. Una Exploración Metacientífica. Universidad Nacional de Quilmes, Bernal

Giribet G, Wheeler WC (1999) On Gaps. Mol Phylogenet Evol 13:132–143

Grant T, Kluge A (2009) Perspective: Parsimony, explanatory power, and dynamic homology testing. Syst Biodivers 7:357–363

Hennig W (1966) Phylogenetic Systematics. University of Illinois Press, Illinois

Herman JL, Challis CJ, Novák Á, et al (2014) Simultaneous Bayesian Estimation of Alignment and Phylogeny under a Joint Model of Protein Sequence and Structure. Mol Biol Evol 31:2251–2266

Japyassú HF, Machado F de A (2010) Coding behavioural data for cladistic analysis: using dynamic homology without parsimony. Cladistics 26:625–642

Jardine N (1967) The Concept of Homology in Biology. Br J Philos Sci 18:125–139

Kitching IJ, Forey PL, Humphries CJ, Williams DM (1998) Cladistics: The Theory and Practice of Parsimony Analysis, 2nd edition. Oxford University Press, Oxford, New York

Kluge AG, Grant T (2006) From conviction to anti-superfluity: old and new justifications of parsimony in phylogenetic inference. Cladistics 22:276–288

Lankester ER (1870) On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. Ann Mag Nat Hist Zool Bot Geol 6:34–43

Martin J, Blackburn D, Wiley EO (2010) Are node-based and stem-based clades equivalent? Insights from graph theory. PLOS Curr Tree Life

Mayr E (1982) The Growth of Biological Thought. Harvard University Press, Cambridge, MA

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Owen R (1848) On the Archetype and Homologies of the Vertebrate Skeleton. John Van Voorst, London

Owen R (1849) On the nature of Limbs. John Van Voorst, London

Patterson C (1982) Morphological characters and homology. In: Joysey KA, Friday AE (eds) Problems of Phylogenetic Reconstruction. Academic Press Inc., London, pp 21–74

Patterson C (1988) Homology in classical and molecular biology. Mol Biol Evol 5:603–625

Pearson CH (2010) Pattern Cladism, Homology, and Theory-Neutrality. Hist Philos Life Sci 32:475–492

Platnick NI (1977) Cladograms, Phylogenetic Trees, and Hypothesis Testing. Syst Biol 26:438–442.

Quinn A (2017) When is a Cladist Not a Cladist? Biol Philos 32:581–598

Ramírez MJ (2007) Homology as a parsimony problem: a dynamic homology approach for morphological data. Cladistics 23:588–612

Remane A (1952) Die Grundlagen des natürlichen Systems der vergleichenden Anatomie und der Phylogenetik. Geest & Portig, Leipzig

Rieppel O (1988) Fundamentals of comparative biology. Birkhäuser Verlag, Basel, Boston, Berlin

Rieppel O, Kearney M (2002) Similarity. Biol J Linn Soc 75:59–82

Rieppel O, Kearney M (2007) The Poverty of Taxonomic Characters. Biol Philos 22:95–113

Robillard T, Legendre F, Desutter-Grandcolas L, Grandcolas P (2006) Phylogenetic analysis and alignment of behavioral sequences by direct optimization. Cladistics 22:602–633

Roffé AJ, Ginnobili S, Blanco D (2018) Theoricity, observation and homology: a response to Pearson. Hist Philos Life Sci 40:42

Saint-Hilaire ÉG (1830) Principes de philosophie zoologique. Pichon et Didier, Paris

Sereno PC (2007) Logical basis for morphological characters in phylogenetics. Cladistics 23:565–587

Simon C, Frati F, Beckenbach A, et al (1994) Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers. Ann Entomol Soc Am 87:651–701

Simpson GG (1961) Principles of Animal Taxonomy. Columbia University Press, New York

Sober E (1988) Reconstructing The Past: Parsimony, Evolution, and Inference. MIT Press, Cambridge, Massachusetts

Vogt L (2018) Towards a semantic approach to numerical tree inference in phylogenetics. Cladistics 34:200–224.

Wheeler WC (1996) Optimization Alignment: The End of Multiple Sequence Alignment in Phylogenetics? Cladistics 12:1–9

Wheeler WC (1999) Fixed Character States and the Optimization of Molecular Sequence Data. Cladistics 15:379–385

Wheeler WC (2001a) Homology and DNA Sequence Data. In: Wagner GP (ed) The Character Concept in Evolutionary Biology. Academic Press, pp 303–317

Wheeler WC (2001b) Homology and the Optimization of DNA Sequence Data. Cladistics 17:S3–S11

Wheeler WC (2002) Optimization Alignment: Down, Up, Error, and Improvements. In: DeSalle R, Giribet G, Wheeler WC (eds) Techniques in Molecular Systematics and Evolution. Birkhäuser Basel, Basel, pp 55–69

Wheeler WC (2003a) Iterative pass optimization of sequence data. Cladistics 19:254–260

Wheeler WC (2003b) Search-based optimization. Cladistics 19:348–355

Wheeler WC (2003c) Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. Cladistics 19:261–268

Wheeler WC (2006) Dynamic homology and the likelihood criterion. Cladistics 22:157–170

Wheeler WC, Aagesen L, Arango CP, et al (2006) Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using Poy. American Museum of Natural History

Wheeler WC, Lucaroni N, Hong L, et al (2015) POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31:189–196

Wiley EO (1979) Cladograms and Phylogenetic Trees. Syst Zool 28:88–92.

Wiley EO, Lieberman BS (2011) Phylogenetics: Theory and Practice of Phylogenetic Systematics. John Wiley & Sons, New Jersey

Xu X, Clark JM, Mo J, et al (2009) A Jurassic ceratosaur from China helps clarify avian digital homologies. Nature 459:940–944.

Young RL, Wagner GP (2011) Why Ontogenetic Homology Criteria Can Be Misleading: Lessons From Digit Identity Transformations. J Exp Zoolog B Mol Dev Evol 316B:165–170.