A laid-back trip through the Hennigian Forests

Evgeny V. Mavrodiev[1], Christopher Dell[1,*] and Laura Schroder[2,3,*]

[1] University of Florida, Florida Museum of Natural History, Gainesville, FL, USA
[2] Department of Anatomy and Neurobiology of University of Tennessee, University of Tennessee Health Science Center Washington, Memphis, TN, USA
[3] Washington, Wyoming, Alaska, Montana and Idaho Medical Education Program, University of Idaho, Moscow, ID, USA
[*] These authors contributed equally to this work.

ABSTRACT

Background. This paper is a comment on the idea of matrix-free Cladistics. Demonstration of this idea's efficiency is a major goal of the study. Within the proposed framework, the ordinary (phenetic) matrix is necessary only as "source" of Hennigian trees, not as a primary subject of the analysis. Switching from the matrix-based thinking to the matrix-free Cladistic approach clearly reveals that optimizations of the character-state changes are related not to the real processes, but to the form of the data representation.

Methods. We focused our study on the binary data. We wrote the simple ruby-based script FORESTER version 1.0 that helps represent a binary matrix as an array of the rooted trees (as a "Hennigian forest"). The binary representations of the genomic (DNA) data have been made by script *1001*. The Average Consensus method as well as the standard Maximum Parsimony (MP) approach has been used to analyze the data.

Principle findings. The binary matrix may be easily re-written as a set of rooted trees (*maximal* relationships). The latter might be analyzed by the Average Consensus method. Paradoxically, this method, if applied to the Hennigian forests, *in principle* can help to identify clades *despite* the absence of the direct evidence from the primary data. Our approach may handle the clock- or non clock-like matrices, as well as the hypothetical, molecular or morphological data.

Discussion. Our proposal clearly differs from the numerous phenetic alignment-free techniques of the construction of the phylogenetic trees. Dealing with the relations, not with the actual "data" also distinguishes our approach from all optimization-based methods, if the optimization is defined as a way to reconstruct the sequences of the character-state changes on a tree, either the standard alignment-based techniques or the "direct" alignment-free procedure. We are not viewing our recent framework as an alternative to the three-taxon statement analysis (3TA), but there are two major differences between our recent proposal and the 3TA, as originally designed and implemented: (1) the 3TA deals with the three-taxon statements or minimal relationships. According to the logic of 3TA, the set of the minimal trees must be established as a binary matrix and used as an input for the parsimony program. In this paper, we operate directly with maximal relationships written just as trees, not as binary matrices, while also using the Average Consensus method instead of the MP analysis. The solely 'reversal'-based groups can always be found by our method without the separate scoring of the putative reversals before analyses.

INTRODUCTION

Here it would seem more appropriate to re-write the characters . . . . in a tree-form representing the relationships exactly

Williams & Ebach (*2006*: 414)

... synapomorphies might evolve more than once. Suppose, indeed, that all synapomorphies really have multiple origins. Would our understanding of relationships then be any different than it now is?

Nelson (*2011*: 140)

Matrix-free Cladistics (*Platnick, 1993*; *Williams & Ebach, 2006*; *Williams & Ebach, 2008*; *Zaraguëta-Bagils et al., 2012*) is a way to represent the binary, morphological and molecular data in tree form, therefore dispensing with the traditional (ordinary) matrix. Demonstration of this idea's efficiency is a major goal of the study.

Recently, the matrix-free Cladistic approach has been developed by *Zaraguëta-Bagils et al. (2012)*. Our framework, however, is different from their proposal, which we discussed below. In general, we are viewing the method developed by *Zaraguëta-Bagils et al. (2012)* as their original conceptualization of the three-taxon statement analysis (3TA) (*Nelson & Platnick, 1991*) which is essentially based on the principle of the maximum compatibility (see *Wilkinson, 1994a* for the initial discussion).

One may note that our vision of the topic replaces the standard "processes-based" subject with a purely geometric view of the problem. Another may see that our viewpoint is closely related to the old discussions around "Transformed cladistics" (*Platnick, 1979*; *Platnick, 1985*) or "pattern-cladistics", or "Cladistics" (with the capital "C", see for example *Hull, 1988*). Numerous works have been published on topic in the past (e.g., *Betty, 1982*; *Brady, 1982*; *De Queiroz & Donoghue, 1990*; *Nelson, 1989*; *Nelson & Platnick, 1981*; *Patterson, 1980*; *Patterson, 1982*; *Platnick, 1979*; *Platnick, 1982*; *Platnick, 1985* see also *Brady, 1985*; *Brower, 2000*; *Eldredge & Cracraft , 1980*; *Ebach & Williams, 2013*; *Nelson, 1972*; for the detailed reviews of the topic see *Hull, 1988*; *Farris, 2014*, *Scott-Ram, 1990* and especially *Williams & Ebach, 2008*). We would like the reader to make his own conclusions regarding the actual summary of these mostly aged but still inspiring debates, which in our minds, were terminated artificially. Of course, it would be impossible to reproduce anything but the shadow of that story in this Introduction.

However, today it is still critical to repeat that dealing with the 'trees' (or with the statements of relationships) can be viewed separately from any considerations of the process (e.g., *Kitching et al., 1998*: p. 1; *Patterson, 1980*: p. 239). Also, the statements of the relationships may be easily treated just *as invariants* of the processes (either real or hypothetical). For example, when considering the statement A(BC) it is easy to conclude

that it does not really matter what the actual cause or process that linked taxa B and C altogether was.

We agree that the consistent implementation of the basic idea of Hennig—grouping solely on synapomorphies returns us to purely comparative thinking (e.g., *Nelson, 2004*; *Williams & Ebach, 2008*), even in the case of molecular data. From that, the treatment of the standard optimization procedures, performed either under the maximum parsimony or other criteria, as something that is related to the reality with its processes, appears to be very naive. In our mind, switching from the matrix-based thinking to the matrix-free Cladistic approach clearly reveals that the optimizations of the character-state changes are related not to the reality *per se*, but to the form of the data representation.

The aspects of the Cladistics we approved are clearly summarized in *Patterson (1980)*:

> "*Hennig's 1966 book, as the title Phylogenetic Systematics suggests, was based in evolutionary theory… But as the theory of cladistics has developed, it has been realized that more and more of the evolutionary framework is inessential, and may be dropped. The chief symptom of this change is the significance attached to nodes in cladograms. In Hennig's book, as in all early work in cladistics, the nodes are taken to represent ancestral species. This assumption has been found to be unnecessary, even misleading, and may be dropped. Platnick (1979) refers to the new theory as "transformed cladistics" and the transformation is away from dependence on evolutionary theory. Indeed, Gareth Nelson, who is chiefly responsible for the transformation, put it like this in a letter to me this summer: "In a way, I think we are merely rediscovering pre-evolutionary systematics, or if not rediscovering it, fleshing it out""*

(*Patterson, 1980*: p. 239; italics ours).

As Nelson himself defined later, Cladistics is "*ordering homologies into a parsimonious hierarchy, which specifies both taxa and their characters (apomorphies)*" (*Nelson, 1989*: 279), and therefore it is independent from the principle of common descent (*Nelson, 1989*).

Among others, *De Queiroz & Donoghue (1990)* have criticized this vision of the topic. In our mind, their most important contra-arguments are:

(1) Nelson's vision of cladistics is strongly related to the "model of nested hierarchy", even if they are supposedly independent of the principle of common descent (*De Queiroz & Donoghue, 1990*: p. 62);

(2) the conventional phylogenetic systematics has greater explanatory power than that which underlies what Nelson calls "cladistics" (*De Queiroz & Donoghue, 1990*: p. 61, see also *Farris et al., 1995* and *Farris, 2013*).

The later criticism of the transformed cladistics has been developed by James Farris and his school (see *Farris et al., 1995* and *Farris, 1997*; *Farris, 2011*; *Farris, 2012*; *Farris, 2014* for the discussions and the summaries of the arguments), essentially in relation to the possible issues of the three-taxon statement analysis (*Nelson & Platnick, 1991*), that we, contrary to *Brower (2000)*, are viewing as the first analytical implementation of the transformed cladistics approach (e.g., *Kitching et al., 1998*). Below we show that the major empirical claims against 3TA, established by Farris and co-authors (see *Siebert & Williams, 1998*,

Mavrodiev, 2015b and *Mavrodiev, 2016* for the reviews), are not relevant to our recent framework.

In our mind, the whole focus of the "explanatory power" (e.g., *De Queiroz & Donoghue, 1990*; *Farris et al., 1995*) is essentially misleading due to its strong relation to the particular philosophy of science that different authors are following or implying. For example, this focus is not quite relevant within the framework of critical philosophy (*Mavrodiev, 2016*, see also *Van de Vijver et al., 2005* and *Kolen & Van de Vijver, 2007* among others), which has likely never been mentioned in the discussed context until recently.

A. Summary of used and implied concepts with selected references and clarifying citations:

a. An ordinary (phenetic) matrix and matrix-free Cladistics (*Platnick, 1993*; *Williams & Ebach, 2006*; *Williams & Ebach, 2008*);

"*The notion that systematic data constitute a normal characters x taxon matrix is not an intrinsically cladistic notion; indeed, that type of matrix seems to have originated with pheneticists. Consider an alternative view, that the three-taxon matrix instead constitute systematic data…*" (*Platnick, 1993*: p. 271; italics ours).

"The data matrix can be generally viewed in three different ways for systematics and biogeography, which we refer to as phylo-phenetics, phylogenetic systematics (transformational) and Cladistic. *Each views the matrix as a combination of manipulating points (taxa, areas and characters) and cells (character states), expressed in a table. All current methods, except Cladistics, treat the transformations or switches between cells and points as the basis for discovering and expressing relationships …for most methods the data matrix is simply a phenetic device for optimising homologues* rather than determining homologies and discovering relationships"
(*Williams & Ebach, 2006*: p. 409; italics ours).

b. The symmetry between binary character and cladogram, as conceptualized by Williams (*1994* and *1996*);

"*The binary character (as represented in a column in a matrix) and the cladogram (as represented by branching diagram) are one and the same thing,* representations of relationships"
(*Williams, 1994*: pp. 451–452, italics ours).

c. Maximal relationship (*Nelson & Platnick, 1981*), summarized in *Williams & Ebach (2006)* and *Williams & Ebach (2008)*, see also *Wilkinson* (*1994b*: p. 344–345).

"…the component (11) is only a "part" of the data (the 00 is a part too), hence to express a specific relationship, both aspects of the data require consideration. Thus, the relationship is AB(CD). *This relationship can be thought of as "maximal" (after (Nelson & Platnick, 1981), in that it includes all the taxic points* (A–D)".
(*Williams & Ebach, 2006*: p. 412, italics ours).

d. Homology as a relationship (summarized in *Williams & Ebach, 2008* and *Nelson, 2011*);

"*For Owen homology is a relation between homologues, not merely the homologues themselves*; and the same for analogy and analogues. For him, the relation between homologues means that the homologues are represented in an archetype, the concept that does the relating (*Williams, 2004*: 196). For vertebrates, his archetype is an ideal (hypothetical) vertebra, or a series of vertebrae and their associated nerves and muscles (cf. *Panchen, 1994*). Hence the relation is that of "similar to" or "same as" an archetype". (*Nelson, 2011*: p. 137, italics ours).

e. Outgroup comparison (*Wiley, 1976*, see also *Platnick & Gertsch, 1976*, *Watrous & Wheeler , 1981* and *Nixon & Carpenter, 1993* for the review);

"*Hennig*'s (*1966*) method differs fundamentally from a purely phenetic method in that all the shared characters are not used to refute a given relationship; only synapomorphous characters are used. Such testing can be accomplished only in an open system, that is, by considering taxa outside the three (or more) taxon system. Such considerations may be termed outgroup comparisons"

(*Wiley, 1976*: p. 11).

"To determine which of two or more homologous states is primitive and which derived, we have used two sources of evidence, immediate outgroup comparison and ontogeny. . . The use of outgroup comparison requires knowledge of the closest relative of the entire group under consideration, which in this case we suggest is the arachnid order Amblypygi. The hypothesis that spiders and amblypygids are sister groups is supported by at least two apparently autapomorphic characters: they are the only arachnids with subchelate chelicerae and with both a pumping pharynx and a pumping stomach . . . We thus hypothesize that any character state found in some but not all spiders and also in amblypygids is plesiomorphic, and its homologs apomorphic; this hypothesis can be falsified in any particular case by incongruence with more numerous synapomorphy patterns. . .

(*Platnick & Gertsch, 1976*: p. 2).

f. *A priori* determination of character's polarity (reviewed in *Kitching et al., 1998* and *Waegele, 2005*; see also *Williams & Ebach, 2008* and *Wiley & Lieberman , 2011*);

"…*when coding polarized characters for cladistics analyses* all plesiomorphic character-states should be marked with a "0" in the data matrix, all apomorphies (when only two character-states are present) with "1"…"

(*Waegele, 2005*: p. 183; italics ours).

g. Grouping solely on synapomorphy (*Hennig, 1966*; see also *Wiley, 1975*, *Wiley, 1976*, *Platnick, 1985*, *Nelson, 2004*, *Williams & Ebach, 2006* and *Williams & Ebach, 2008* among others);

"That a common stem form is shared by a group of species (a condition for a monophyletic group) *can be proven only by means of synapomorphous characters, not by symplesiomorphous characters* "

(*Hennig, 1966*: p. 90; italics ours).

h. Cladistics is more than Wagner's algorithm or its optimization-based derivatives (*Nelson, 2004*; *Williams & Ebach, 2006*; *Williams & Ebach, 2008*);

"For Hennig, "…*we can never directly observe the phylogenetic transformation of a character*". (Summarized by *Nelson, 2004*: p. 134; italics ours).

i. Average Consensus method (*Lapointe & Cucumel, 1997*; *Lapointe & Levasseur, 2004*) and its extensions (*Creevey, 2004*, see also *Lapointe & Levasseur, 2004*);

"The average consensus method combines the information from multiple trees by calculating the path length from every taxa to every other taxa on each of the source trees. This method utilizes branch lengths (if present) or assumes a branch length of unity (a branch length of one) if not present …The specifics of the method are as follows:

- The distance from each taxon to every other taxon on a tree is calculated. This is the equivalent of the sum of the branch lengths in the path between the two taxa.
- The average distance of each taxa to every other taxa across all the source trees is calculated…
- A least squares method is used to estimate a supetree phylogeny (with branch lengths) that best describes the distance matrix …You also have the option of doing a neighbor joining supertree from the calculated distance matrix" (*Creevey, 2004*: p. 18).

j. Matrix representation with Parsimony (*Baum, 1992*; *Ragan, 1992*);

"Rooted phylogenetic trees can be represented as matrices in which the rows correspond to termini, and columns correspond to internal nodes (elements of the *n*-tree). Parsimony analysis of such a matrix will fully recover the topology of the original tree" (*Ragan, 1992*: p. 53).

k. Three-taxon (item) statement analysis (3TA) (*Nelson & Platnick, 1991*); see also *Kitching et al., 1998* and *Williams & Ebach, 2008* for the reviews.

"Three-item statement analysis. A method of cladistic analysis that focuses on the smallest unit of relationship, the three-item statement, rather than on characters. The observed features of taxa are coded in terms of the relationships they imply, that is three-item statements, and the optimal cladogram is that which maximizes the number of accommodated three-item statements" (*Kitching et al., 1998*: p. 218).

B. Initial propositions

A binary character is a tree with one informative node (*Platnick et al., 1996*; *Williams, 1994*; *Williams & Ebach, 2006*; *Williams & Ebach, 2008*). For example, if state 1 is apomorphic, then the character ABCDE/00011 *is a rooted tree* ABC(DE), where (DE) is a clade (monophyletic group) based on the apomorphic character-state. If all of the clades of the tree are based on apomorphic character-states, we call this tree a "Hennigian" tree. If we accept that the trees such as ABC(DE) or A(BC(DE)) are Hennigian, than the trees A((BC)(DE)) or ((ABC)(DE)) are not. The non-Hennigian trees contain groups based

on plesiomorphic characters-states—for example, tree ((ABC)(DE)) is non-Hennigian because group (ABC) based on the plesiomorphic characters-state zero. Hennigian trees may also be easily seen as simple hierarchies of two character-states.

Let A be defined as the outgroup. In this case the *relationship* ABC(DE) may be re-written as A(BC(DE)) (or as a (A(BC(DE)))), *even if, strictly speaking, we do not have any formal evidence for the groups* (BCDE) (or (ABCDE))v (see *Rieppel, 1991*: p. 95 and *Williams & Ebach, 2009* for the related discussions). For example, the *minimal relationship* A(BC)/011 may be re-written as (A(BC)), even if there is no evidence for the group (ABC).

Consider the tree A(BC(DE)) for one more time. If the value of taxon B is missing, two solutions appear to be possible.

First, character ABCDE can be re-written as two trees, assuming that the missing value may be either zero or one: ABC(DE) = A(BC(DE)) or A(C(BDE)). Another possibility implies the exclusion of the taxon B from the tree. This reduces the character ABCDE to the tree AC(DE).

Characters like ABCDE/00001, 000?1, ??001, 00000 etc. collapse to polytomies such as (ABCDE), (ABCE) etc.

The basic idea of Matrix Representation with Parsimony (*Baum, 1992*; *Ragan, 1992*) states that the tree (cladogram) can be represented as a binary matrix. Here we propose something opposite—we accept that the binary character may be re-written as a tree (*Nelson & Ladiges, 1992*, 492–493; *Williams, 1994*; *Siebert & Williams, 1998*: 342; *Williams & Ebach, 2006*: 414; see also *Platnick et al., 1996*). Therefore, the binary matrix can be viewed as a set (forest, array etc.) of branching diagrams (cladograms) that may be called a "Hennigian forest" if one of the character-states (for example, the character-state zero) *a priori* is defined as plesiomorphic.

It is easy to assume that the Hennigian forest might be analyzed at least by some consensus methods, typically treated as a ways of estimating the supertrees (*Gordon, 1986*, see also *Wilkinson, Cotton & Thorley, 2004* and *Bininda-Emonds, 2014* for the reviews) for example, by the Average Consensus method (*Lapointe & Cucumel, 1997*; *Lapointe & Levasseur, 2004*, see also *Bininda-Emonds, 2014*). Using this method as an example, below we are attempting to display the workings of this assumption.

As summarized by *Lapointe & Levasseur* (*2004*: 87), the Average Consensus procedure is a method that takes as input a profile of weighted trees (i.e., trees with the branch lengths *Lapointe & Cucumel, 1997*; *Creevey, 2004*; *Lapointe & Levasseur, 2004*) and returns a consensus tree that is, in some sense, "closest" to the entire profile. Originally this method was designed to combine the clock-like or "ultrametric" rooted trees (or the trees with the total branch length from the root up to any tip equal (*Felsenstein, 2004*: p. 161)), but later have been extended to allow for the combination of all types of the weighted trees, the clock-like or not (reviewed in *Lapointe & Levasseur, 2004*: 88 and *Creevey, 2004*: p. 18).

This method allows us to operate with the Hennigian trees directly, completely excluding the binary data matrix from the analysis. The ordinary (phenetic) matrix remains necessary only as a "source" of the forest of the Hennigian trees, not as a primary subject of the analysis. Within this framework, the criteria of the best trees, such as the minimal number

of character-state changes, as well as the standard optimization procedures, if optimization is defined as a way to reconstruct the sequences of the character-state changes on a tree, all appeared to be unrequired.

The matrix-free Cladistic approach, as designed and implemented by *Zaraguëta-Bagils et al. (2012)* and others (reviewed in *Zaraguëta-Bagils et al., 2012*), also treats the characters as hierarchies or as rooted trees, but clearly differs from our proposal in several aspects. According to the logic and implementation of *Zaraguëta-Bagils et al. (2012)*, the rooted character-state trees must first be reduced to a set of three-item statements. The software LisBeth (*Zaraguëta-Bagils et al., 2012*) performs the modified three-taxon statement analysis, which is based on the maximum congruence of the statements and finally calculates the intersection that the tree is built from, and only from, the three-taxon statements that are common to all the optimal trees and characters (*Zaraguëta-Bagils et al., 2012*).

Our approach also differs *in principle* from the numerous purely phenetic alignment-free techniques of the construction of the phylogenetic trees (reviewed and summarized in *Warnow, 2014* and *Bogusz & Whelan, 2016*) because these methods do not even refer to the concepts of the "relation" and "homology", but operate by computing similarities (the pairwise distances) between the raw (or non-aligned) sequences (e.g., *Warnow, 2014*).

Finally, one may ask why the data cannot be imputed directly as trees (*Zaraguëta-Bagils et al., 2012*) or, in other words, by just completely skipping the data matrix? In principle, it is possible. However so far this solution seemed for us to be impractical in many cases.

MATERIALS AND METHODS

Despite the plethora of publications within the field of contemporary Bioinformatics, *there is no software available to rewrite the binary matrix as a Hennigian forest*. Given this, we wrote the simple ruby-based script FORESTER version 1.0 (named "FORESTER" below) that helps represent the binary data as Hennigian trees for future manipulations (Figs. 1–9).

FORESTER (deposited on https://github.com/dellch/forester) processes each input file by storing the text of each line of the file as an element of an array. It finds the beginning and ending location (index) of the characters in the matrix, and loops from the starting index to the ending index, writing a line of the new file(s) at the end of each loop.

For the binary Matrix, the usage of the script is: **ruby trees.rb –inputfilename**

The pre-defined out-group taxon should be placed last in the matrix before running of the FORESTER.

Three output tree-files are available as a result of the run: the first file contains polytomies such as (ABCDE)(the output tree-file named as a *"With poly…"*), while the second and the third outputs appear without such polytomies, but trees may be rooted relative to *a priori*-defined outgroup taxon (e.g., (A(BC(DE)))) (the output array appears as named as *"No poly…"* tree-file) or to the basal polytomy (e.g., (ABC(DE))) (the output tree-file named as an *"Additional…"* tree-file).

All tree files and the input binaries should be written in the "relaxed" non-interleaved Newick (PHYLIP) format (reviewed in *Felsenstein, 1989*, see also *Maddison & Maddison, 2011* and *Swofford, 2002*).

Most parsimonious tree/average consensus

Polarized binary matrix (Nelson and Platnick (1991): modified Matrix 17)

Characters of the binary matrix (B.), re-written as trees in Newick (Phylip) notation (Hennigian forest)

```
                  O                      A  0000              (O(A(BCDEF)))
               A                         B  1001              (O(AB(CDEF)))
             B                   ←       C  1101      →       (O(ABC(DEF)))
           C                Maximum      D  1111   FORESTER   (O(AEF(BCD)))
         D                  Parsimony    E  1110
                            analysis     F  1110
       E                    (e.g., PAUP*) O  0000
A.     F                            B.                        C.
```

↑ Average consensus Method - step 2 (e.g., PAUP*): calculation of the consensus (A.) based on the Table (D.)

Average consensus Method - step 1 (CLANN): calculation of the the Table (D.) based on the Hennigian forest (C.) ↓

```
O
A  2.000000
B  2.500000  2.500000
C  2.750000  2.750000  2.250000
D  3.000000  3.000000  2.500000  2.250000
E  2.750000  2.750000  2.750000  2.500000  2.250000
F  2.750000  2.750000  2.750000  2.500000  2.250000  2.000000

                        D.
```

Figure 1 (A) Most parsimonious tree of the length is equal to five (CI = 0.80, RI = 0.86) based on the Matrix 17 (B) from *Nelson & Platnick* (*1991*: p. 362) with character four excluded, tree was a posteriori rooted relatively taxon O\the average consensus tree of the score 0.22222 calculated based on the table (D); (B) Matrix 17 from *Nelson & Platnick* (*1991*: p. 362) with the character four excluded; (C) four remaining characters of the Matrix 17 (B) established as Hennigian trees; (D) the average consensus table based on the forest (C).
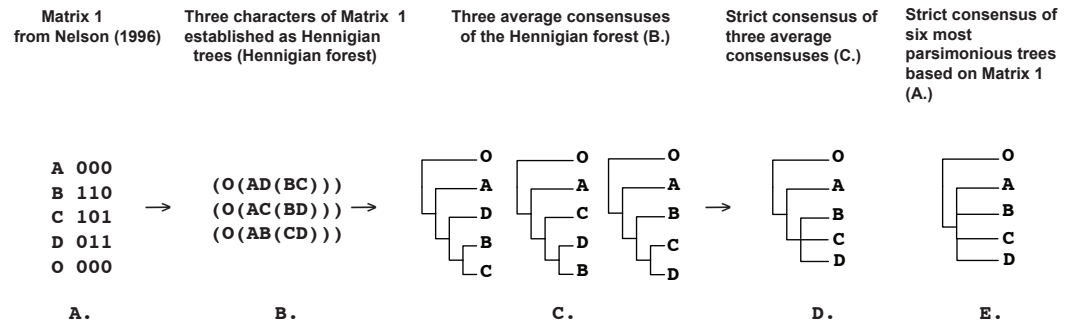
The minimal trees (the three-taxon statements (3TS)) are not the subjects of our recent considerations, but FORESTER contains options for speedy rewriting of the 3TS matrices as the arrays of the minimal trees. The usage should be: **ruby seedlings.rb –inputfilename** with only one output Newick file saved as a result of the run.

Seven binary matrices (Figs. 1–4, 5, 9) are taken from the literature (*Nelson, 1996*; *Kluge, 1994*; *Farris, 1997*; *Farris & Kluge, 1998*; *Kluge & Farris, 1999*).

Preparing the DNA supermatrix of the multiple loci of the Great Apes (Hominidae, Primates, Mammalia) (Fig. 6) we first used MUSCLE (*Edgar, 2004*) to align separate 98 gene regions listed in the modified summary of the *Lehtonen et al. (2011)*. The resulting separate DNA alignments were concatenated in the SeaView 4.0 (*Gouy, Guindon & Gascuel, 2010*) after removing purely aligned regions using the Gblocks 0.91 (*Talavera & Castresana, 2007*) as implemented in a Seaview (*Gouy, Guindon & Gascuel, 2010*). The Gblocks less stringent selection strategy (*Talavera & Castresana , 2007*) was decided to be the best choice. Eventually, the only characters with non-ambiguous value of the outgroup (*Papio*) have been saved within concatenated 42,222 bp alignment.

Matrix 1 from Nelson (1996)	Three characters of Matrix 1 established as Hennigian trees (Hennigian forest)	Three average consensuses of the Hennigian forest (B.)	Strict consensus of three average consensuses (C.)	Strict consensus of six most parsimonious trees based on Matrix 1 (A.)

```
A 000
B 110        (O(AD(BC)))
C 101   →    (O(AC(BD)))  →
D 011        (O(AB(CD)))
O 000

A.              B.              C.              D.              E.
```

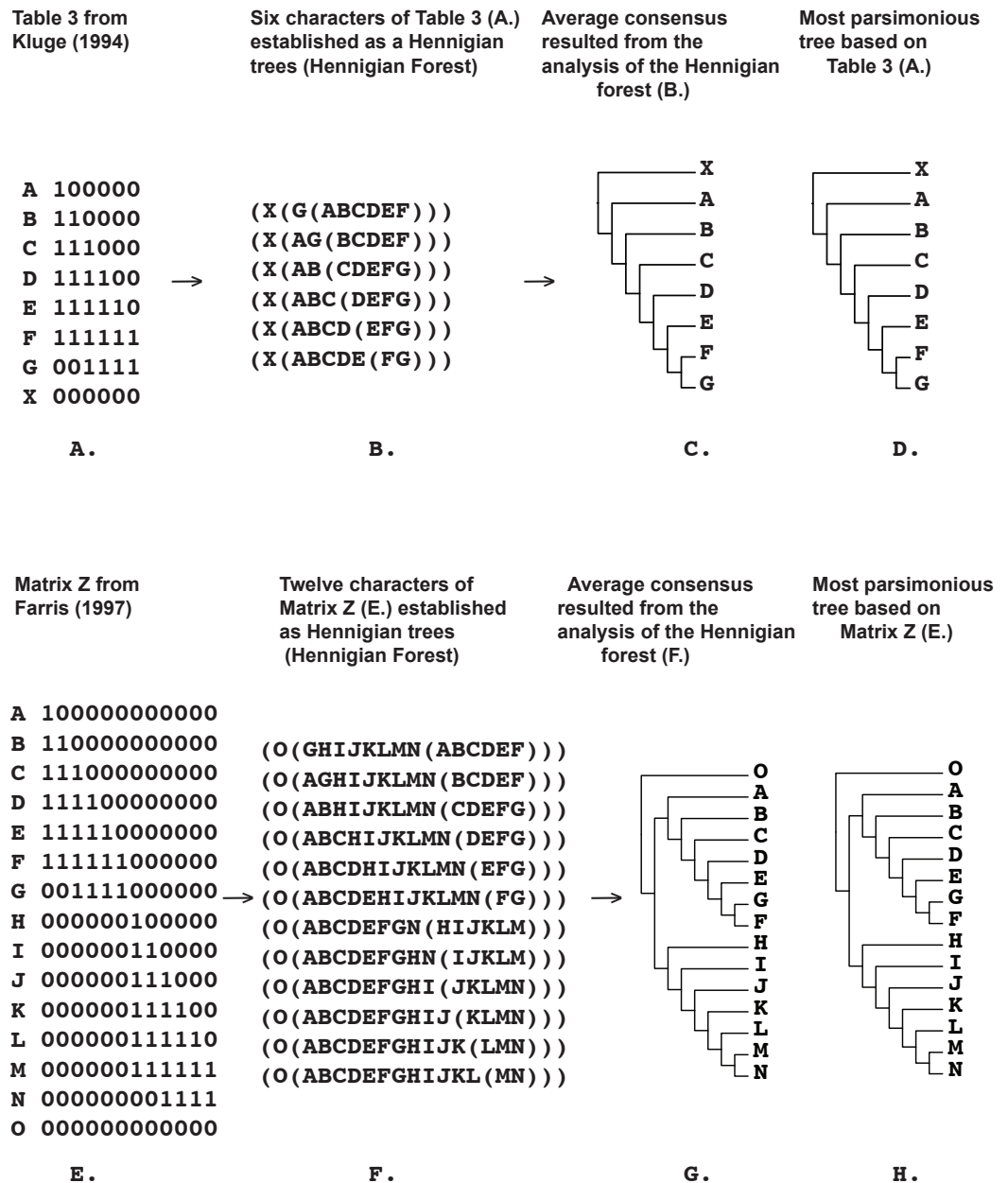**Figure 2** (A) Matrix 1 modified from *Nelson* (*1996*: p. 1) (*Nelson, 1996*; *Williams & Ebach, 2005*); (B) three conflicting characters of Matrix 1 from *Nelson* (*1996*: p. 1) (A) (see also *Williams & Ebach, 2005*) established as a Hennigian trees; (C) three average consensuses of the forest (B) of the score zero and their strict consensus (D); (E) strict consensus of six most parsimonious trees of the length equal to five (CI = 0.6000, RI = 0.3333) based on the Matrix 1 from *Nelson* (*1996*: p. 1). All six trees were *a posteriori* rooted relatively taxon O.

Table 2 from Kluge (1994)	Nine characters of the Table 2 (A.) established as Hennigian trees (Hennigian Forest)	Strict consensus of three average consensuses resulted from the analysis of the Hennigian forest (B.)	Most parsimonious tree based on Table 2 (A.)

```
A 100000000
B 110000000     (X(IJK(ABCDEFGH)))
C 111000000     (X(AJK(BCDEFGHI)))
D 111100000     (X(AB(CDEFGHIJK)))
E 111110000     (X(ABC(DEFGHIJK)))
F 111111111  →  (X(ABCD(EFGHIJK)))  →
G 111111111     (X(ABCDE(FGHIJK)))
H 111111111     (X(ABCDE(FGHIJK)))
I 011111111     (X(ABCDE(FGHIJK)))
J 001111111     (X(ABCDE(FGHIJK)))
K 001111111
X 000000000

A.              B.              C.              D.
```

Figure 3 (A) Table 2 modified from *Kluge* (*1994*: p. 408); (B) nine characters of the Table 2 from *Kluge* (*1994*: p. 408) (A) established as a Hennigian trees; (C) strict consensus of three average consensuses of the score 0.27118 resulted the analysis of the forest (B); (D) most parsimonious tree of the length = 11 (CI = 0.8182, RI = 0.9429) based on the Table 2 from *Kluge* (*1994*: p. 408). Tree was *a posteriori* rooted relatively taxon X.

Matrix of 166 muscular characters of Primates (Mammalia) and *Rattus* (Mammalia, Rodentia, Muridae) has been modified from the image of *Diogo & Wood* (*2012*: p. 121, Table 3.1). We treated all multistate characters from Table 3.1 (*Diogo & Wood* (*2012*: p. 121) as ordered (reviewed in *Kitching et al., 1998*) (Fig. 7) and recoded all of them in additive binary form (reviewed in *Kitching et al., 1998*) using TAXODIUM ver. 1.2 (*Mavrodiev & Madorsky, 2012*).

Figure 5 from Farris & Kluge (1998)

Twenty-three characters of Figure 5 (A.) established as Hennigian trees (Hennigian Forest)

Average consensus resulted from the analysis of the Hennigian forest (B.)

Most parsimonious tree based on Figure 5 (A.)

```
O 0000000000000000000000000
A 0000000000000000000000000
B 1000000000000000000000000
C 1100000000000000000000000
D 1110000000000000000000000
E 1111000000000000000000000
F 1111100000000000000000000
G 1111110000000000000000000
H 1111111000000000000000000
I 1111111100000000000000000
J 1111111110000000000000000
K 1111111111000000000000000
L 1111111111100000000000000
M 1111111111110000000000000
N 1111111111111000000000000
P 1111111111111100000000000
Q 1111111111111110000000000
R 1111111111111111000000000
S 1111111111111111100000000
T 1111111111111111110000000
U 1111111111111111111111111
V 0011111111111111111111111
W 0001111111111111111111111
X 0000111111111111111111111
Y 0000011111111111111111111
Z 0000011111111111111111111
```

```
(O(AVWXYZ(BCDEFGHIJKLMNPQRSTU)))
(O(ABVWXYZ(CDEFGHIJKLMNPQRSTU)))
(O(ABCWXYZ(DEFGHIJKLMNPQRSTUV)))
(O(ABCDXYZ(EFGHIJKLMNPQRSTUVW)))
(O(ABCDEYZ(FGHIJKLMNPQRSTUVWX)))
(O(ABCDEF(GHIJKLMNPQRSTUVWXYZ)))
(O(ABCDEFG(HIJKLMNPQRSTUVWXYZ)))
(O(ABCDEFGH(IJKLMNPQRSTUVWXYZ)))
(O(ABCDEFGHI(JKLMNPQRSTUVWXYZ)))
(O(ABCDEFGHIJ(KLMNPQRSTUVWXYZ)))
(O(ABCDEFGHIJK(LMNPQRSTUVWXYZ)))
(O(ABCDEFGHIJKL(MNPQRSTUVWXYZ)))
(O(ABCDEFGHIJKLM(NPQRSTUVWXYZ)))
(O(ABCDEFGHIJKLMN(PQRSTUVWXYZ)))
(O(ABCDEFGHIJKLMNP(QRSTUVWXYZ)))
(O(ABCDEFGHIJKLMNPQ(RSTUVWXYZ)))
(O(ABCDEFGHIJKLMNPQR(STUVWXYZ)))
(O(ABCDEFGHIJKLMNPQRS(TUVWXYZ)))
(O(ABCDEFGHIJKLMNPQRST(UVWXYZ)))
(O(ABCDEFGHIJKLMNPQRST(UVWXYZ)))
(O(ABCDEFGHIJKLMNPQRST(UVWXYZ)))
(O(ABCDEFGHIJKLMNPQRST(UVWXYZ)))
(O(ABCDEFGHIJKLMNPQRST(UVWXYZ)))
```


A.   B.   C.   D.

Figure 4 (A) Figure 5 modified from *Farris & Kluge* (*1998*: p. 353); (B) 23 characters of Figure 5 modified from *Farris & Kluge* (*1998*: p. 353) (A) established as a Hennigian trees; (C) the average consensus tree of the score 1.00184 resulted the analysis of the forest (B); (D) most parsimonious tree of the length = 28 (CI = 0.8214, RI = 0.9711) based on the Table 2 from *Farris & Kluge* (*1998*: p. 353). Tree was *a posteriori* rooted relatively taxon O.

The 15,9074 bp alignment of complete plastomes of palms (Arecales), Dasypogonales, and *Typha latifolia* (Poales, Typhaceae) (Fig. 8) has been downloaded from the online supplement of *Barrett et al.* (*2016*: their Fig. 4 and Fig. S5).

The binary representations of the genomic (DNA) data (Figs. 6, 8 and 9) have been made by script *1001* (*Mavrodiev, 2015a*). In theory, different binary representations of the DNA alignments (*Mavrodiev, 2015a*) can be chosen as sources for the Hennigian trees. But in this paper, for future analyses, we decided to select the simplest binary matrices that resulted the "presence–absence" recoding (reviewed in *Kitching et al. 1998*) of the DNA alignments (*Mavrodiev, 2015a*), but saving the only characters that corresponded to the value of zero of the outgroup taxon. Script *1001* (*Mavrodiev, 2015a*) is the easiest way to generate such binary matrix. However, it can be prepared even manually using the text editor.

The Average Consensus tables are calculated following the default settings of Clann version 4.1.5 (http://chriscreevey.github.io/clann/) (*Creevey, 2004*; *Creevey & McInerney, 2005*; *Creevey & McInerney, 2009*) for the **avcon** command with the branch lengths of the

Table 3 from Kluge (1994)

```
A 100000
B 110000
C 111000
D 111100    →
E 111110
F 111111
G 001111
X 000000
```

A.

Six characters of Table 3 (A.) established as a Hennigian trees (Hennigian Forest)

```
(X(G(ABCDEF)))
(X(AG(BCDEF)))
(X(AB(CDEFG)))
(X(ABC(DEFG)))
(X(ABCD(EFG)))
(X(ABCDE(FG)))
```

B.

Average consensus resulted from the analysis of the Hennigian forest (B.)

→

C.

Most parsimonious tree based on Table 3 (A.)

D.

Matrix Z from Farris (1997)

```
A 100000000000
B 110000000000
C 111000000000
D 111100000000
E 111110000000
F 111111000000
G 001111000000 →
H 000000100000
I 000000110000
J 000000111000
K 000000111100
L 000000111110
M 000000111111
N 000000001111
O 000000000000
```

E.

Twelve characters of Matrix Z (E.) established as Hennigian trees (Hennigian Forest)

```
(O(GHIJKLMN(ABCDEF)))
(O(AGHIJKLMN(BCDEF)))
(O(ABHIJKLMN(CDEFG)))
(O(ABCHIJKLMN(DEFG)))
(O(ABCDHIJKLMN(EFG)))
(O(ABCDEHIJKLMN(FG)))
(O(ABCDEFGN(HIJKLM)))
(O(ABCDEFGHN(IJKLM)))
(O(ABCDEFGHI(JKLMN)))
(O(ABCDEFGHIJ(KLMN)))
(O(ABCDEFGHIJK(LMN)))
(O(ABCDEFGHIJKL(MN)))
```

F.

Average consensus resulted from the analysis of the Hennigian forest (F.)

→

G.

Most parsimonious tree based on Matrix Z (E.)

H.

Figure 5 (A) Table 3 modified from *Kluge* (*1994*: p. 409); (B) six characters of Table 3 from *Kluge* (*1994*: p. 409) (A) established as a Hennigian trees; (C) the average consensus tree of the score 0.29213 resulted the analysis of the forest (B); (D) most parsimonious tree of the length is equal to eight (CI = 0.7500, RI = 0.8182) based on the Table 3 from *Kluge* (*1994*: p. 408). Tree was *a posteriori* rooted relatively taxon X; (E) Matrix Z from *Farris* (*1997*: p. 136); (F) 12 characters of Matrix Z from *Farris* (*1997*: p. 136) (A) established as a Hennigian trees; (G) the average consensus tree of the score 1.9603e−08 resulted the analysis of the forest (F); (H) most parsimonious tree of the length = 16 (CI = 0.7500, RI = 0.8947) based of Matrix Z from *Farris* (*1997*: p. 136). Tree was *a posteriori* rooted relatively taxon O.

Most parsimonious resulted from standard MP analysis of 42222 bp 98-*loci* DNA sumermatrix of the Great Apes based on the modified summary of *Lehtonen et al. (2011)*

The average consensus resulted from the analysis of the Hennigian forests of the binary representations of 142222 bp 98-*loci* DNA sumermatrix of the Great Apes based on the modified summary of *Lehtonen et al. (2011)* (A)

The same as (B), but with all 168 trees, which contain the clade (*Homo* plus *Pan*) have been removed from the input forest (B)

Figure 6 (A) Tree resulted the standard MP analysis of the 98-*loci* supermatrix of the Great Apes (Hominidae, Primates, Mammalia) constructed using the modified summary of *Lehtonen et al. (2011)* (see "Materials and Methods" for the details) and *a posteriori* rooted relatively *Papio* (length = 12,092, CI = 0.7877, RI = 0.5597); 34,022 characters are constant, and the number of the parsimony-informative characters is equal to 4,311. MP BS values are provided below branches. The only characters with non-ambiguous value of the outgroup (*Papio*) have been saved within concatenated alignment; (B) the average consensus tree of the score 0.00084 resulted the analysis of the Hennigian forest of the 5,507 trees derived from the binary representation of the 98-*loci* supermatrix of the Great Apes (Hominidae, Primates, Mammalia) (see "Materials and Methods" for the details); *Papio* is assumed to be the best all-plesiomorphic group. "Additional…" FORESTER's output tree file (see "Materials and Methods" for the details) selected for future analyses; (C) the average consensus tree of the score 0.00090 resulted the analysis of the Hennigian forest of the 5,339 trees derived from the modified binary representation of the 98-*loci* supermatrix of the Great Apes (Hominidae, Primates, Mammalia) (see 6.I. and "Materials and Methods" for the details), but with all 168 trees, which contain the clade (Homo plus Pan) have been removed from the input forest. "Additional…" FORESTER's output tree file (see "Materials and Methods") selected for future analyses.

all input trees assigned to unity (*Creevey, 2004*) and also used as future inputs for PAUP* 4.0b10 (*Swofford, 2002*) (PAUP* hereafter). All input-trees are treated as weighted equally.

In most of the analyses the optimality criterion of the best tree is defined in the simplest way—as a distance with non-weighted least squares (*Creevey, 2004*; *Creevey & McInerney, 2005*; *Creevey & McInerney, 2009*; *Swofford, 2002*). In the case of the analysis of the Hennigian forest of Matrix Z (*Farris, 1997*) (Fig. 5E–5H) we used the weighted least squares (exponential weights with $P' = 7$ (*Swofford, 2002*)). Different formal criteria, such as "Balanced Minimal evolution" (BME) (*Desper & Gascuel, 2002*; *Swofford, 2002*) (Figs. 7B and 7C) can also be used.

All standard Maximum Parsimony (MP) analyses as well as Maximum Parsimony Bootstrap (MP BS) estimation (Fig. 6) have been conducted in PAUP* (*Swofford, 2002*), sometimes (Figs. 7 and 8) as implemented in CIPRESS (*Miller, Pfeiffer & Schwartz, 2010*)

Most parsimonious tree resulted from standard MP analysis of the Table 3.1 modified from *Diogo & Wood (2012)* (166 muscular characters)

The average consensus resulted from the analysis of the Hennigian forests of the Table 3.1 modified from *Diogo & Wood (2012)* (A) and recoded in additive binary form

The same as (B), but with single tree, which contain the clade (*Homo* plus *Pan*) has been removed from the input forest (B)
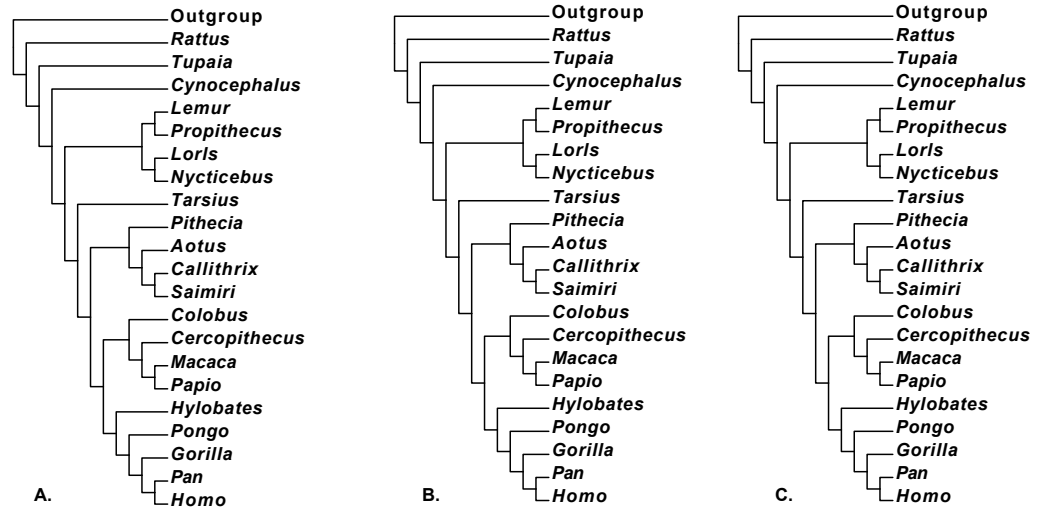
Figure 7 (A) Tree resulted standard MP analysis of modified Table 3.1 with 166 muscular characters of Primates (Mammalia) and *Rattus* (Mammalia, Rodentia, Muridae) from *Diogo & Wood* (*2012*: p. 121), *a posteriori* rooted relatively added artificial all-zero taxon (length = 307, CI = 0.5700, RI = 0.7295). 50 characters are parsimony-uninformative, and the number of the parsimony-informative characters is equal to 116. All multistate characters (60, 68, 124, 129, 136, 138, 149, and 162) have been treated as ordered; (B) the average consensus tree of the score 23.83826 resulted the analysis of the Hennigian forest of the 123 trees derived from the Table 3.1 modified from *Diogo & Wood* (*2012*: p. 121) with all multistate characters recoded in additive binary form; (C) the average consensus tree of the score 23.84628 resulted the analysis of the Hennigian forest of the 122 trees derived from complete Table 3.1 modified from *Diogo & Wood* (*2012*: p. 121) with all multistate characters recoded in additive binary form, but with the single tree with the clade (*Homo* plus *Pan*) has been removed from the input forest.

using 1,000 random addition replicates (saving no more than 100 trees per replicate), and with the TBR branch swapping/MulTrees option in effect; the gaps were treated as "missing entities".

The routine manipulations with the matrices and the tree-files were performed with Mesquite v. 3.01 (*Maddison & Maddison, 2011*), PAUP* (*Swofford, 2002*) and FigTree v. 1.4.2 (*Rambaut, 2012*).

RESULTS AND DISCUSSION

In Cladistics, clades must be based solely on "derived" or apomorphic character-states (e.g., *Hennig, 1966*; *Platnick, 1985*; *Nelson, 2004*; *Williams & Ebach, 2008*). For example, in the case of the binary matrix, where the character state "zero" is defined as "plesiomorphic" before the analysis, all clades should be based solely on the state "one" (e.g., *Platnick, 1985*). However, today the 3TA (*Nelson & Platnick, 1991*) is the only method, which completely avoids grouping on plesiomorphy (e.g., *Williams & Ebach, 2008*). Therefore, analyzing all of the selected matrices (Figs. 1–9) we have followed the general logic of 3TA. Specifically,

Most parsimonious tree resulted from standard MP analysis of 159074 bp alignment of complete plastomes of palms (Arecales), Dasypogonales, and *Typha latifolia* (Poales, Typhaceae) from *Barrett et al. (2016)*

Two identical average consensuses resulted from the analysis of the Hennigian forests of the binary representations of 159074 bp alignment of complete plastomes of palms (Arecales), Dasypogonales, and *Typha latifolia* (Poales, Typhaceae) (A). In the case of the consensus (C) all 54 trees, which contain the clade (*Cocos nucifera* plus *Elaeis guineensis*) have been removed from the input forest

A.

B., C.
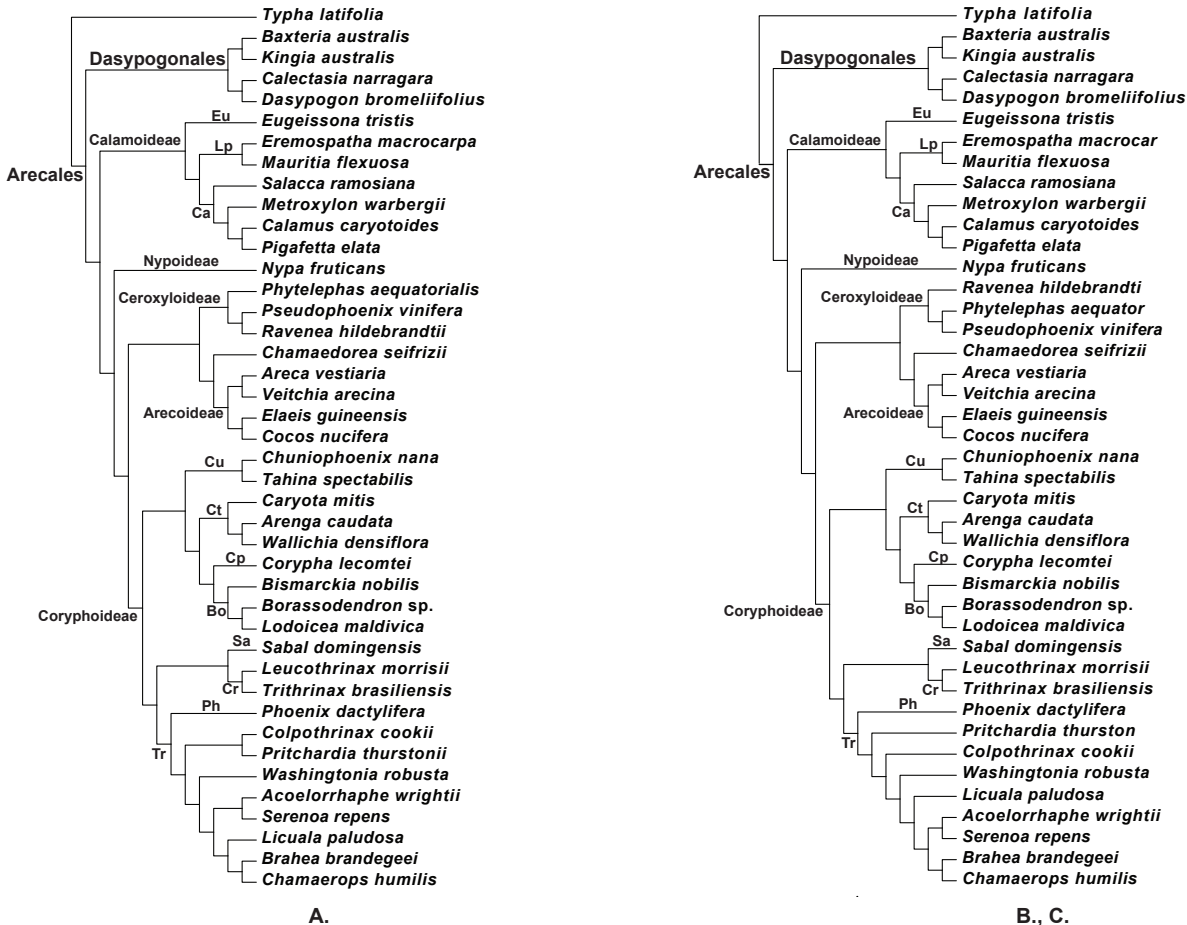
Figure 8 (A) Tree resulted standard MP analysis of 159,074 bp alignment of complete plastomes of palms (Arecales), Dasypogonales, and *Typha latifolia* (Poales, Typhaceae) from *Barrett et al. (2016)* (see their Fig. 4 and Fig. S5), a posteriori rooted relatively the cattail (*Typha*) (length = 66,788, CI = 0.7363, RI = 0.6109), 119,251 characters are constant, and the number of the parsimonyinformative characters is equal to 19,124; (B) the average consensus tree of the score 0.00755 resulted the analysis of the Hennigian forest of the 25,630 trees derived from the modified binary representation of the original 159,074 bp alignment of complete plastomes of palms (Arecales), Dasypogonales, and *Typha latifolia* (Poales, Typhaceae) from *Barrett et al. (2016)* (see "Materials and Methods" for the details). The cattail is assumed to be the best all-plesiomorphic group. "Additional…" FORESTER's output tree file (see "Materials and Methods" for the details) selected for future analysis; (C) the average consensus tree of the score 0.00760 resulted the analysis of the Hennigian forest of the 25,576 trees, derived from the binary representation of the original 15,9074 bp alignment of the complete plastomes of palms (Arecales), Dasypogonales, and *Typha latifolia* (Poales, Typhaceae) from *Barrett et al. (2016)* (see "Materials and Methods" for the details), but with all of the 54 trees, which contain the clade (*Cocos nucifera* plus *Elaeis guineensis*) have been removed from the input forest before analysis.

(1) we tried to explicate all possible Hennigian trees *a priori* to the analyses in order (2) to find the best-fitting trees as the next step.

Five out of the seven binary matrices-examples (Figs. 1, 3, 4, 5) we analyzed have been established as a main sources of the major empirical claim against 3TA (and in our mind, against the whole idea of grouping solely on synapomorphies)—the principle inability

Figure 9 (A) Hypothetical DNA matrix from *Kluge & Farris* (*1999*: p. 206); (B) tree resulted standard MP analysis of hypothetical DNA matrix from *Kluge & Farris* (*1999*: p. 206) (length = 33, CI = RI = 1.0000); (C) binary recoding of hypothetical DNA matrix from *Kluge & Farris* (*1999*: p. 206)) (A); (D) tree resulted standard MP analysis of the binary matrix (C) (length = 66, CI = RI = 1.0000); (E) reduced binary matrix (C) with only characters that corresponded to the value of zero of the outgroup taxon (O) have been saved; (F) Hennigian Forest of the reduced binary matrix (E). "Additional…" FORESTER's output tree file (see "Materials and Methods" for the details) selected for future analysis; (G) tree resulted standard MP analysis of the reduced binary matrix (E) (length = 33, CI = RI = 1.0000); tree was a posteriori rooted relatively taxon O\average consensus of the Hennigian Forest (F) (score is equal to zero); (H) UPGMA phenogram of the hypothetical DNA matrix (A) (see also *Kluge & Farris*, *1999*: p. 206).

of this method to operate with the putative reversals (*Kluge, 1994*; *Farris, 1997*; *Farris & Kluge, 1998* reviewed in *Siebert & Williams, 1998*, *Mavrodiev, 2015b* and *Mavrodiev, 2016*). Using these critical examples we are demonstrated (Figs. 1, 3, 4 and 5) that this claim is not relevant to our recent proposal. Also using the original critical example of *Kluge & Farris (1999)* we also showed that our method could successfully handle the non-clock molecular data (Fig. 9).

We are not viewing our recent approach as an alternative to 3TA. However, there are two major differences between our recent proposal and the 3TA as originally designed and implemented by *Nelson & Platnick (1991)*. (1) The 3TA deals with the three-taxon statements (3TS) or *minimal* relationships (*Williams & Siebert, 2000*; *Williams & Ebach, 2006*; *Williams & Ebach, 2008*). The set of 3TSs must be established as a binary matrix and used as an input for the parsimony program (*Nelson & Platnick, 1991*; see also *Williams & Siebert, 2000*, *Williams & Ebach, 2006*. *Williams & Ebach, 2008* and *Mavrodiev & Madorsky, 2012*).

In this paper, we operate directly with *maximal* relationships (*Nelson & Platnick, 1981*; summarized in *Williams & Ebach, 2006* and *Williams & Ebach, 2008*) *written just as trees*, not as binary matrices, while also using the Average Consensus method instead of the MP analysis.

Dealing with relations, not with actual "data", clearly distinguishes our approach from all optimization-based methods, both the standard alignment-based methods and the "direct" alignment-free approaches (e. g., *Wheeler, 1996*; *Wheeler, 2001*).

In summary:

(1) Script FORESTER helps to rewrite every character of the binary matrix as a tree in a Cladistics way basing all of the groups solely on *a priori* defined apomorphic character-state "1".

(2) Next, the forest of these "maximal" Hennigian trees can be used to calculate their average consensus (*Lapointe & Cucumel, 1997*; *Lapointe & Levasseur, 2004*) (Figs. 1–9).

A possible issue of the Hennigian approach to the data is the inability to operate with the putative reversals (reviewed in *Farris & Kluge, 1998*, *Siebert & Williams, 1998* and *Mavrodiev (2015b)*, unless, however, the 'reversals' are scored as separate apomorphic character-states before the analysis (*Mavrodiev, 2015b*).

Paradoxically, the Average Consensus method, if applied to the Hennigian forests, helps to identify the 'reversal'-based clades without the separate scoring of the putative reversals, or, in other words, *despite* the absence of evidence from the primary data (Figs. 1, 3 and 4). A similar effect had been described by *Nelson & Platnick (1991)* for 3TA and discussed in more detail by *Siebert & Williams (1998)* as well as by *Mavrodiev (2016)* who named this paradox as a "Synthetic Theorem of Nelson and Platnick".

For example, consider the modified Matrix 17 from *Nelson & Platnick (1991)* (Fig. 1). According to the MP optimization, group (EF) is based on the plesiomorphic character-state zero and therefore may be treated by someone as a putative 'reversal'-based clade (e.g., *Farris, 1997*). However as explained above, grouping based on the plesiomorphic character state is prohibited within the Hennigian framework (e.g., *Platnick, 1985*; *Nelson, 2004*; *Williams & Ebach, 2008*). Therefore the standard MP solution for the group (EF) is clearly non-Hennigian or phenetic (e.g., *Williams & Ebach, 2008*; *Mavrodiev, 2015b*).

None of the trees from the array c. (Fig. 1) *contains any plesiomorphic-based groups, but clade (EF) is still defined in the average consensus tree* (Fig. 1) *throughout the analysis of forest c.* (Fig. 1) *or, in other words, despite the lack of evidence from the primary data* (see also *Nelson & Platnick, 1991*: 363 for the similar discussion). The same situation is detected in the cases of Table 1 from *Kluge (1994)* (Fig. 3), and Table 5 from *Kluge & Farris (1999)* (Fig. 4).

The Average Consensus technique, if applying to the Hennigian forests, may also help to avoid more potentially negative effects of putative reversals. For example, regarding his Table 3, *Kluge (1994*: 408–410) mentioned that taxa G and F are highly supported sisters with taxon G exhibiting reversals only in characters one and two. As discovered by *Kluge (1994)*, the 3TA (*Nelson & Platnick, 1991*) removed taxon G from F, despite the strong evidence of their relationship. A similar situation appears in the case of Matrix Z, designed

by *Farris (1997)* simply by duplication of the Table 3 from *Kluge (1994)* (*Farris, 1997*; *Siebert & Williams, 1998*).

However, the tree-shapes of the average consensuses of the Hennigian forests of the Table 3 from *Kluge (1994)* (Fig. 5A–5D) and Matrix Z from *Farris (1997*: 136) (Fig. 5E–5H) appeared to be identical to the topologies, which were the result of the MP analyses of Table 3 and Matrix Z (*Kluge, 1994*; *Farris, 1997*). 3TA, with fractional weighting procedure (*Nelson & Ladiges, 1992*; *Williams & Ebach, 2008*) can also compensate the negative effect of putative reversals in the same situations (*Siebert & Williams, 1998*).

Like the 3TA (*Nelson & Platnick, 1991*), the Average Consensus analysis of the forests of maximal relationships can successfully recognize groups for which the standard optimization criteria of the MP analysis produce no unequivocal synapomorphies (*Nelson, 1996*; *Williams & Ebach, 2005*) (Fig. 2). For example, group (BCD) is successfully recognizable after the Average Consensus analysis of the three trees, each representing the conflicting binary characters from Matrix 1 from *Nelson (1996)* (see also *Williams& Ebach, 2005*; *Williams & Ebach, 2006*) (Fig. 2).

Williams & Ebach (*2006*: 414) offered a very similar solution for the clade (BCD). These authors mentioned that *it would seem more appropriate to re-write the characters of Matrix 1 from* Nelson (1996) *in a tree-form representing the relationships exactly*, such that the three "characters" are AD(BC), AC(BD) and AB(CD), which, when "*combined*" (*Williams & Ebach, 2006*: 414, italics are ours), unambiguously provide evidence for the solution A(BCD) (*Williams & Ebach, 2006*: 414). In contrast to this intuitively clear solution, as well as to the results of the Average Consensus analysis (Fig. 2), neither of strict, majority rule, Adams', or combinable component consensuses (see *Kitching et al., 1998* and *Swofford, 2002* for the reviews and implementation of the methods) of the three trees AD(BC), AC(BD), and AB(CD) (*Williams & Ebach, 2006*: 414) are able to recognize the group (BCD).

Williams & Ebach (*2006*: 414) are also noted that the re-writing of the binary characters effectively converts an ordinary (phenetic) matrix into a "Cladistic matrix". In our mind, a Cladistic matrix is similar, but still not of the same entity as a simplest forest of the Hennigian trees (see *Siebert & Williams, 1998*: 242 for the initial elementary examples of the Cladistic matrices called by these authors as a "tabular formulations of the 3TA" of the conventional data; see also *Williams & Ebach, 2006*: 412, *Nelson & Ladiges, 1992*: 492–493 and *Williams & Ebach, 2008*).

In the first approach, the simplest options of the search (the least-square criteria of the fit plus un-weighted algorithms) should be sufficient to estimate the reasonable average consensuses. However, the Average Consensus approach allows various schemes of weighing as well as different optimization criteria (*Creevey, 2004*; *Lapointe & Levasseur, 2004*). For example, the use of the exponential weights of the least squares as implemented in PAUP* (*Swofford, 2002*) or BME-algorithm (*Desper & Gascuel, 2002*; see (*Swofford, 2002* for recent implementation) may increase the efficiency of the search (Figs. 5 and 7).

The proposed approach may also handle the clock-like (e.g., Fig. 6) or non clock-like matrices (e.g., Fig. 9), as well as the hypothetical or the real data, either the DNA sequence data or the morphological characters (Figs. 6–8). There are several possible ways to

do this. As a tentative test, for someone it may seem to be necessary to re-recode the conventional multistate matrices as a binary matrices, to establish these matrices as a Hennigian forests, and to estimate the average consensus trees that fit these Hennigian forests (Figs. 6–8). Therefore, even in the cases of the big molecular data (Fig. 8), the taxon-character matrices (molecular alignments) may be necessary only as sources of the arrays of maximal relationships.

We also would like to stress, that for the real data it is also true that at least in some cases the clades may appear within the final tree without the direct evidence from the primary data. For example, the clades (*Homo* plus *Pan*) or (*Cocos nucifera* plus *Elaeis guineensis*) are still recognizable by the Average Consensus method, even if all of the trees from the corresponding Hennigian arrays, that contain the same clades (*Cocos nucifera* plus *Elaeis guineensis*) or (*Homo* plus *Pan*), will be excluded from the forest-inputs (Figs. 6C, 7C and 8C). In other words, the clades (*Homo* plus *Pan*) or (*Cocos nucifera* plus *Elaeis guineensis*) appears within the final average consensus trees (Figs. 6–8) as a kind of *relations* (e.g., *Nelson, 2011*), rather than the character-based entities.

CONCLUSIONS

1. Binary matrix with *a priori* defined plesiomorphic character-state may be re-written as a simple set of rooted branching diagrams (as a "Hennigian forest").
2. The last might be analyzed by the Average Consensus method.
3. This procedure eventually avoids the taxon-character matrix from the analysis of the data. Within the proposed framework, the ordinary (phenetic) matrix is necessary only as a "source" of the Hennigian trees, not as a primary subject of the analysis.
4. Within this approach, the criteria of the best trees based on the character-state changes are not required.
5. Dealing with the relations, not with the actual "data", clearly distinguishes our approach from all optimization-based techniques, either the standard taxon-character alignment-based methods, or the "direct" alignment-free method.
6. The Average Consensus method, if applied to the Hennigian forests, may helps to identify the clades *despite* the absence of the direct evidence from the primary data. Because of such ability, the 'reversal'-based clades can always be found by this approach without the separate scoring of the putative reversals.

grouping. Finally, we appreciate the important notes on the earliest, the pre-submission version of the text that have been made by the editorial staff of *Acta Biotheoretica*. No agreement with methodology, general context and the conclusions of this paper is implied from the behalf of any person listed in above.

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Evgeny V. Mavrodiev conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Christopher Dell contributed reagents/materials/analysis tools.
- Laura Schroder wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:
GitHub: https://github.com/dellch/forester.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3578#supplemental-information.

REFERENCES

Barrett CF, Baker WJ, Comer JR, Conran JG, Lahmeyer SC, Leebens-Mack JH, Li J, Lim GS, Mayfield-Jones DR, Perez L, Medina J, Pires JC, Santos C, Stevenson DWM, Zomlefer WB, Davis JI. 2016. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *The New Phytologist* **209**:855–870 DOI 10.1111/nph.13617.

Mavrodiev et al. (2017), *PeerJ*, DOI 10.7717/peerj.3578

20/25

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**:3–10 DOI 10.2307/1222480.

Betty J. 1982. Classes and cladists. *Systematic Zoology* **31**:25–34 DOI 10.2307/2413411.

Bininda-Emonds OR. 2014. An introduction to supertree construction (and partitioned phylogenetic analyses) with a view toward the distinction between gene trees and species trees. In: L Garamszegi Z, ed. *Modern phylogenetic comparative methods and their application in evolutionary biology*. Berlin-Heidelberg: Springer, 49–76.

Bogusz M, Whelan S. 2016. Phylogenetic tree estimation with and without alignment: new distance methods and benchmarking. *Systematic Biology* **66**:218–231 DOI 10.1093/sysbio/syw074.

Brady RH. 1982. Theoretical issues and ''Pattern cladistics''. *Systematic Zoology* **31**:286–291 DOI 10.2307/2413235.

Brady RH. 1985. On the independence of systematics. *Cladistics* **1**:113–126 DOI 10.1111/j.1096-0031.1985.tb00416.x.

Brower AVZ. 2000. Evolution is not a necessary assumption of cladistics. *Cladistics* **16**:143–154 DOI 10.1111/j.1096-0031.2000.tb00351.x.

Creevey C. 2004. Clan: construction of supertrees and exploration of phylogenomic information from partially overlapping datasets (version 3.0.0), user manual. The lab of James McInerney, National University of Ireland, Galway. *Available at https://pdfs.semanticscholar.org/b32d/8462fa952bdd2d588fdb3e96e3f694cea891.pdf* .

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**:390–392 DOI 10.1093/bioinformatics/bti020.

Creevey CJ, McInerney JO. 2009. Trees from trees: construction of phylogenetic supertrees using clann. *Bioinformatics for DNA Sequence Analysis* **537**:139–161 DOI 10.1007/978-1-59745-251-9_7.

De Queiroz K, Donoghue MJ. 1990. Phylogenetic systematics or Nelson's version of cladistics? *Cladistics* **6**:61–75 DOI 10.1111/j.1096-0031.1990.tb00525.x.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. *Journal of Computational Biology* **9**:687–705 DOI 10.1089/106652702761034136.

Diogo R, Wood BA. 2012. *Comparative anatomy and phylogeny of primate muscles and human evolution*. Boca Raton: CRC Press, Taylor & Francis Group, LLC.

Ebach MC, Williams DM, Vanderlaan TA. 2013. Implementation as theory, hierarchy as transformation, homology as synapomorphy. *Zootaxa* **3641**:587–594 DOI 10.11646/zootaxa.3641.5.7.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797 DOI 10.1093/nar/gkh340.

Eldredge N, Cracraft J. 1980. Phylogenetic patterns and the evolutionary process. New York: Columbia University Press.

Farris JS. 1997. Cycles. *Cladistics* **13**:131–144.

Farris JS. 2011. Systemic foundering. *Cladistics* **27**:207–221 DOI 10.1111/j.1096-0031.2010.00331.x.

Farris JS. 2012. 3ta sleeps with the fishes. *Cladistics* **28**:422–436
DOI 10.1111/j.1096-0031.2011.00390.x.

Farris JS. 2013. Symplesiomorphies and explanation. *Cladistics* **29**:13–14
DOI 10.1111/j.1096-0031.2012.00434.x.

Farris JS. 2014. Pattern poses. *Cladistics* **30**:116–119 DOI 10.1111/cla.12034.

Farris JS, Kallersjo M, Albert VA, Allard M, Anderberg A, Bowditch B, Bult C, Carpenter JM, Crowe TM, DeLaet J, Fitzhugh K, Frost D, Goloboff P, Humphries CJ, Jondelius U, Judd D, Karis PO, Lipscomb D, Luckow M, Mindell D, Muona J, Nixon K, Presch W, Seberg O, Siddall ME, Struwe L, Tehler A, Wenzel J, Wheeler Q, Wheeler W. 1995. Explanation. *Cladistics* **11**:211–218
DOI 10.1111/j.1096-0031.1995.tb00086.x.

Farris JS, Kluge AG. 1998. A/the brief history of three-taxon analysis. *Cladistics* **14**:349–362 DOI 10.1111/j.1096-0031.1998.tb00343.x.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 32). *Cladistics* **5**:164–166.

Felsenstein J. 2004. *Inferring phylogenies.* Sunderland: Sinauer Associates Inc.

Gordon AD. 1986. Consensus supertrees—the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* **3**:335–348
DOI 10.1007/BF01894195.

Gouy M, Guindon S, Gascuel O. 2010. Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**:221–224 DOI 10.1093/molbev/msp259.

Hennig W. 1966. *Phylogenetic systematics.* Urbana: University of Illinois Press.

Hull D. 1988. *Science as a process: an evolutionary account of the social and conceptual development of science.* Chicago: University of Chicago Press.

Kitching IJ, Forey PL, Humphries CJ, Williams DM. 1998. Cladistics. The theory and practice of parsimony analysis. In: *Systematics association publication, 11.* 2nd edition. Oxford: Oxford University Press.

Kluge AG. 1994. Moving targets and shell games. *Cladistics* **10**:403–413
DOI 10.1111/j.1096-0031.1994.tb00186.x.

Kluge A, Farris JS. 1999. Taxic homology = overall similarity. *Cladistics* **15**:205–212.

Kolen F, Van de Vijver G. 2007. Philosophy of biology: naturalistic or transcendental? *Acta Biotheoretica* **55**:35–46 DOI 10.1007/s10441-007-9009-6.

Lapointe FJ, Cucumel G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology* **46**:306–312 DOI 10.1093/sysbio/46.2.306.

Lapointe FJ, Levasseur C. 2004. Everything you always wanted to know about the average consensus, and more. In: Bininda-Emonds RP, ed. *Phylogenetic supertrees: combing information to reveal the Tree of Life. Computational biology.* vol. 3. Dordrecht: Springer, 87–105.

Lehtonen S, Saaksjarvi IE, Ruokolainen K, Tuomisto H. 2011. Who is the closest extant cousin of humans? Total-evidence approach to hominid phylogenetics via simultaneous optimization. *Journal of Biogeography* **38**:805–808 DOI 10.1111/j.1365-2699.2010.02354.x.

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 3.01. *Available at http://mesquiteproject.org/*.

Mavrodiev EV. 2015a. *1001*—a tool for binary representations of unordered multistate characters (with examples from genomic data). *PeerJ PrePrints* **3**:e1403 DOI 10.7287/peerj.preprints.1153v1.

Mavrodiev EV. 2015b. Three-taxon analysis can always successfully recognize groups based on putative reversals. *PeerJ PrePrints* **3**:e1206 DOI 10.7287/peerj.preprints.979v1.

Mavrodiev EV. 2016. Dealing with propositions, not with the characters: the ability of three-taxon statement analysis to recognize groups based solely on 'reversals', under the maximum-likelihood criteria. *Australian Systematic Botany* **29**:119–125 DOI 10.1071/SB16006.

Mavrodiev EV, Madorsky A. 2012. TAXODIUM version 10: a simple way to generate uniform and fractionally weighted three-item matrices from various kinds of biological data. *PLOS ONE* **7(11)**:e48813 DOI 10.1371/journal.pone.0048813.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In: Pirece M, ed. *'2010 gateway computing environments workshop (GCE)', 14 November 2010, Proceedings CFP1022F-POD*. New Orleans: Institute of Electrical and Electronics Engineers, 45–52.

Nelson GJ. 1972. Phylogenetic relationship and classification. *Systematic Zoology* **21**:227–231 DOI 10.2307/2412293.

Nelson G. 1989. Cladistics and evolutionary models. *Cladistics* **5**:275–289 DOI 10.1111/j.1096-0031.1989.tb00490.x.

Nelson G. 1996. *Nullius in verba*. New York: Self-published.

Nelson G. 2004. Cladistics: its arrested development. In: Williams DM, Forey PL, eds. *Milestones in systematics, Systematics Association Special Volume Series 67*. Boca Raton: CRC Press, 127–147.

Nelson G. 2011. Resemblance as evidence of ancestry. *Zootaxa* **2946**:137–141.

Nelson G, Ladiges PY. 1992. Information-content and fractional weight of 3-item statements. *Systematic Biology* **41**:490–494 DOI 10.1093/sysbio/41.4.490.

Nelson G, Platnick NI. 1981. *Systematics and biogeography: cladistics and Vicariance*. New York: Columbia University Press.

Nelson G, Platnick NI. 1991. Three-taxon statements—a more precise use of parsimony? *Cladistics* **7**:351–366 DOI 10.1111/j.1096-0031.1991.tb00044.x.

Nixon KC, Carpenter JM. 1993. On outgroups. *Cladistics* **9**:413–426 DOI 10.1111/j.1096-0031.1993.tb00234.x.

Panchen AL. 1994. Richard Owen and the concept of homology. In: Hall BK, ed. *Homology: the hierarchical basis of comparative biology*. San Diego: Academic Press, 21–62.

Patterson C. 1980. Cladistics. *Biologist* **27**:234–240.

Patterson C. 1982. Classes and cladists or individuals and evolution. *Systematic Zoology* **31**:284–286 DOI 10.2307/2413234.

Platnick NI, Gertsch WJ. 1976. The suborders of spiders: a cladistic analysis (Arachnida, Araneae). *American Museum Novitates* **2607**:1–15.

Platnick NI. 1979. Philosophy and the transformation of cladistic. *Systematic Zoology* **28**:537–546 DOI 10.2307/2412566.

Platnick NI. 1982. Defining characters and evolutionary groups. *Systematic Zoology* **31**:282–284 DOI 10.2307/2413233.

Platnick NI. 1985. Philosophy and the transformation of cladistics revisited. *Cladistics* **1**:87–94 DOI 10.1111/j.1096-0031.1985.tb00412.x.

Platnick NI. 1993. Character optimization and weighting—differences between the standard and three-taxon approaches to phylogenetic inference. *Cladistics* **9**:267–272 DOI 10.1111/j.1096-0031.1993.tb00224.x.

Platnick NI, Humphries CJ, Nelson G, Williams DM. 1996. Is Farris optimization perfect?: three-taxon statements and multiple branching. *Cladistics* **12**:243–252 DOI 10.1111/j.1096-0031.1996.tb00011.x.

Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* **1**:53–58 DOI 10.1016/1055-7903(92)90035-F.

Rambaut A. 2012. FigTree v. 1.4.2 Molecular evolution, phylogenetics and epidemiology. Edinburgh: University of Edinburgh, Institute of Evolutionary Biology.

Rieppel O. 1991. Things, taxa and relationships. *Cladistics* **7**:93–100 DOI 10.1111/j.1096-0031.1991.tb00024.x.

Scott-Ram NR. 1990. *Transformed cladistics, taxonomy and evolution.* Cambridge: Cambridge University Press.

Siebert DJ, Williams DM. 1998. Recycled. *Cladistics* **14**:339–347 DOI 10.1111/j.1096-0031.1998.tb00342.x.

Swofford DL. 2002. *PAUP*. Phylogenetic analysis using parsimony (*and other methods).* Sunderland: Sinauer Associates.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**:564–577 DOI 10.1080/10635150701472164.

Van de Vijver G, Van Speybroeck L, De Waele D, Kolen F, De Preester H. 2005. Philosophy of biology: outline of a transcendental project. *Acta Biotheoretica* **53**:57–75 DOI 10.1007/s10441-005-5350-9.

Waegele JW. 2005. *Foundations of phylogenetic systematics.* München: Pfeil Verlag.

Warnow T. 2014. Computational phylogenetics. An introduction to designing methods for phylogeny estimation. *Available at http://tandy.cs.illinois.edu/textbook.pdf* .

Watrous LE, Wheeler QD. 1981. The out-group comparison method of character analysis. *Systematic Zoology* **30**:1–11 DOI 10.2307/2992297.

Wheeler W. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* **12**:1–9 DOI 10.1111/j.1096-0031.1996.tb00189.x.

Wheeler W. 2001. Homology and the optimization of DNA sequence data. *Cladistics* **17**:S3–S11 DOI 10.1111/j.1096-0031.2001.tb00100.x.

Wiley EO. 1975. Popper, Kr, systematics, and classification—reply to Bock, W and other evolutionary taxonomists. *Systematic Zoology* **24**:233–243 DOI 10.2307/2412764.

Wiley EO. 1976. The phylogeny and biogeography of fossil and recent gars (Actinopterygii: Lepisosteidae). *Miscellaneous Publication—University of Kansas, Museum of Natural History* **64**:1–111.

Wiley EO, Lieberman BS. 2011. *Phylogenetics: the theory and practice of phylogenetic systematics*. 2nd edition. Hoboken: John Wiley & Sons, Inc.

Wilkinson M. 1994a. Three-taxon statements—when is a parsimony analysis also a clique analysis? *Cladistics* **10**:221–223 DOI 10.1111/j.1096-0031.1994.tb00174.x.

Wilkinson M. 1994b. Common cladistic information and its consensus representation—reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* **43**:343–368 DOI 10.1093/sysbio/43.3.343.

Wilkinson M, Cotton JA, Thorley JL. 2004. The information content of trees and their matrix representations. *Systematic Biology* **53**:989–1001 DOI 10.1080/10635150490522737.

Williams DM. 1994. Combining trees and combining data. *Taxon* **43**:449–453 DOI 10.2307/1222721.

Williams DM. 1996. Characters and cladograms. *Taxon* **45**:275–283 DOI 10.2307/1224668.

Williams DM. 2004. Homologues and homology, phenetics and cladistics: 150 years of progress. In: Williams DM, Forey PL, eds. *Milestones in systematics. Systematics association special volume series*, vol. 67. Boca Raton: CRC Press, 181–224.

Williams DM, Ebach MC. 2005. Drowning by numbers: rereading Nelson's "Nullius in verba". *Botanical Review* **71**:415–447 DOI 10.1663/0006-8101(2005)071[0415:DBNRNN]2.0.CO;2.

Williams DM, Ebach MC. 2006. The data matrix. *Geodiversitas* **28**:409–420.

Williams DM, Ebach MC. 2008. *Foundations of systematics and biogeography*. New York: Springer.

Williams DM, Ebach MC. 2009. What, exactly, is cladistics? Re-writing the history of systematics and biogeography. *Acta Biotheoretica* **57**:249–268 DOI 10.1007/s10441-008-9058-5.

Williams DM, Siebert DJ. 2000. Characters, homology and three-item analysis. In: Scotland RW, Pennington RT, eds. *Homology and systematics: coding characters for phylogenetic analysis. Systematics Association Special Volume Series 58*, London: Chapman and Hall, 183–208.

Zaraguëta-Bagils R, Ung V, Grand A, Vignes-Lebbe R, Cao N, Ducasse J. 2012. LisBeth: new cladistics for phylogenetics and biogeography. *Comptes Rendus Palevol* **11**:563–566 DOI 10.1016/j.crpv.2012.07.002.