

Essays on the Binary Representations of the DNA Data

Evgeny V. Mavrodiev ^{1,*}  and Nicholas E. Mavrodiev ²

¹ Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

² Santa Fe College, 3000 NW 83rd St, Gainesville, FL 32606, USA; nickmavrick11@gmail.com

* Correspondence: evgeny@ufl.edu

Abstract: The advancement of modern genomics has led to the large-scale industrial production of molecular data and scientific outcomes. Simultaneously, conventional DNA character alignments (sequence alignments) are utilized for DNA-based phylogenetic analyses without further recoding procedures or any *a priori* determination of character polarity, contrary to the requirements of foundations of phylogenetic systematics. These factors are the primary reasons why the binary perspective has not been implemented in modern molecular phylogenetics. In this study, we demonstrate how to recode conventional DNA data into various types of binary matrices, either unpolarized or with established polarity. Despite its historical foundation, our analytical approach to DNA sequence data has not been adequately explored since the inception of the molecular age. Binary representations of conventional DNA alignments allow for the analysis of molecular data from a purely comparative or static perspective. Furthermore, we show that the binarization of DNA data possesses broad mathematical and cultural connotations, making them intriguing regardless of their applications to different phylogenetic procedures.

Keywords: DNA; binary characters; character polarity; phylogenetic systematics; cladistics; Leibniz; High Chinese Classics

1. Introduction

Advancements in modern genomics have facilitated the large-scale production of molecular data and scientific papers. Typically, this ‘industrial’ production results in publications by large teams of co-authors who adhere to a limited number of methods and concepts [1]. For many, this situation represents unprecedented progress in scientific knowledge. However, it is unsurprising that the same case prompts some to look back to understand what might have been overlooked as the scientific community transitioned into its current state of an ‘industrial enterprise’ over the last couple of decades. It appears that such individuals constitute a minority within the scientific community, and the first author of this text is among them. In other words, this article is not everyone’s cup of coffee. It consists of two essays that may seem unrelated. The first is devoted to the binary representation of DNA data in the context of phylogenetic analyses. The second addresses selected philosophical and related issues concerning the binary recoding of DNA data, primarily referencing Gottfried Wilhelm Leibniz (1646–1716), an outstanding German philosopher and mathematician of the 17th and 18th centuries. We leave it to the reader to make the connection between these two parts of the paper. The latter, therefore, aims to briefly outline some perspectives that have been missed or ignored by both modern phylogenetics and current structural studies of the DNA molecule. Both essays are semantically independent, and the second cannot be considered a philosophical conclusion of the first. However, both essays argue in favor of extending the philosophical and methodological understanding of



Academic Editor: Carol A Stepien

Received: 12 December 2024

Revised: 22 January 2025

Accepted: 7 February 2025

Published: 14 February 2025

Citation: Mavrodiev, E.V.; Mavrodiev, N.E. Essays on the Binary Representations of the DNA Data. *DNA* **2025**, *5*, 10. <https://doi.org/10.3390/dna5010010>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

what the DNA molecule is. The match of such argumentation seems to be the warranty against the impression that there is no connection between the two essays of this study.

Since the time of Leibniz, representing information in binary form has been a valuable practice. However, it remains unclear why DNA sequence data cannot be translated into a binary format within a phylogenetic context. This translation, if analytically adjusted, is the primary focus of the first essay. We discuss the potential application of the binary representation of DNA sequence data to phylogenetic analysis and the connotations of this application. In the second essay, we argue that the binary representation of DNA data has broad mathematical and cultural undertones and is engaging regardless of its applications to phylogenetic analyses.

Although binary representations of conventional DNA data can be successfully analyzed within the framework of parametric phylogenetics, we have chosen to omit the relevant discussion and examples here, considering that the logical issues surrounding the maximum likelihood methodology remain unresolved [2]. We believe that molecular phylogenetics is more than modern parametric approaches that operate with conventional molecular alignments, and DNA is more than a combination of the four basic symbols, even within the formal phylogenetic framework. To some, the latter proposition may sound trivial, but below, we try to demonstrate that this is not necessarily the case.

2. Binary Representation of DNA Sequence Data and Molecular Phylogenetics

The title of Willi Hennig's monograph is *Phylogenetic Systematics* [3]. For simplicity, we refer to any method related to Hennig's methodology (Hennig, 1966) as "phylogenetic analysis". In other words, in this study, we use the term "phylogenetic" narrowly, meaning it pertains to the basic principles of Hennig's study [3]. Consequently, phylogenetic analyses include pattern-cladistics [4], conventional cladistics [5], and parametric approaches [6,7]. These analytical methods can all be technically termed "phylogenetic analyses", although the strict relation of the latter two to the phylogenetic systematics of Willi Hennig is debatable [2,4].

The term 'binary representation' is a homonym: any alignment of multistate characters can be represented in binary form either as it is or through comparison with the assumed outgroup taxon. However, the straightforward issue of comparative binary representations of DNA data, and consequently the issue of polarity, has not been adequately addressed or properly discussed in molecular systematics.

The explanation of some basic terms within practical dimensions is necessary. The digital polarized binary matrix is a Nexus of Phylip (see below) file that consists of a table comprising two numerical symbols: zero and one, as well as a question mark. The meaning of these symbols follows Hennigian principles and is defined *a priori* to analysis [3,4]—the symbol "one" denotes an "apomorphic" ("derived") character, the symbol "zero" signifies a "plesiomorphic" ("primitive") character, and the question mark indicates ambiguity. In an unpolarized binary matrix, the states zero and one do not represent a hypothesis of character polarity. In the simplest digital "presence-absence" binary matrix, the character state "zero" represents the absence of a particular nucleotide, while the character state "one" indicates its presence. When the character state "zero" ("absence") is assumed to be plesiomorphic, the interpretation of the presence-absence matrix aligns with that of the polarized binary matrix.

In this study, the binary matrix is a digital representation of the standard alignment of DNA characters or the matrix consisting of the five symbols: A, T, C, G, and a question mark (assuming ambiguities such as R = A + G are disregarded). The meanings of these symbols are straightforward; for instance, the symbol "A" represents "adenine". In this context, the

standard DNA-based alignment does not involve the concept of character–state polarity. For example, before the analysis, whether the character state A (adenine) is apomorphic or plesiomorphic is unknown. Figures 1–4 graphically explain the digitization of DNA characters and, therefore, the polarization of the latter.

Because in cladistics groups must be defined based only on the synapomorphies [3,4,8–11], it is critical to assume states' polarity before analysis and group only on the states "1" of the polarized binary matrix [4,9–11]. Consequently, a fundamental issue in contemporary molecular systematics is the lack of polarization in molecular matrices [9–11], rendering them analytically uninformative from a cladistic perspective [4,12]. Therefore, it is critical to convert the standard DNA alignment to the polarized digital table (matrix) (Figures 1–4). Such a task was in part addressed in the past: three-taxon representations of unordered multistate data, such as conventional alignments of DNA characters, are the straightforward cladistic examples of polarized binary data (see below).

Historically, polarized binary matrices were proposed as an ideal data format for phylogenetic (cladistic) analysis following the argumentation schemes of Willi Hennig, who determined each character's polarity before the construction of a cladogram [4,10–18]. Hennigian logic may be clear even from a purely methodological standpoint: without a character hypothesis in place *a priori* to analysis, we are unable to test hypotheses *a posteriori* [15].

Thus, in an effort to implement Hennigian principles in molecular systematics, the polarized binary matrix can be practically utilized in molecular phylogenetics as an alternative to conventional non-polarized DNA alignments. The illustration of this possibility is the primary goal of the first essay of this paper.

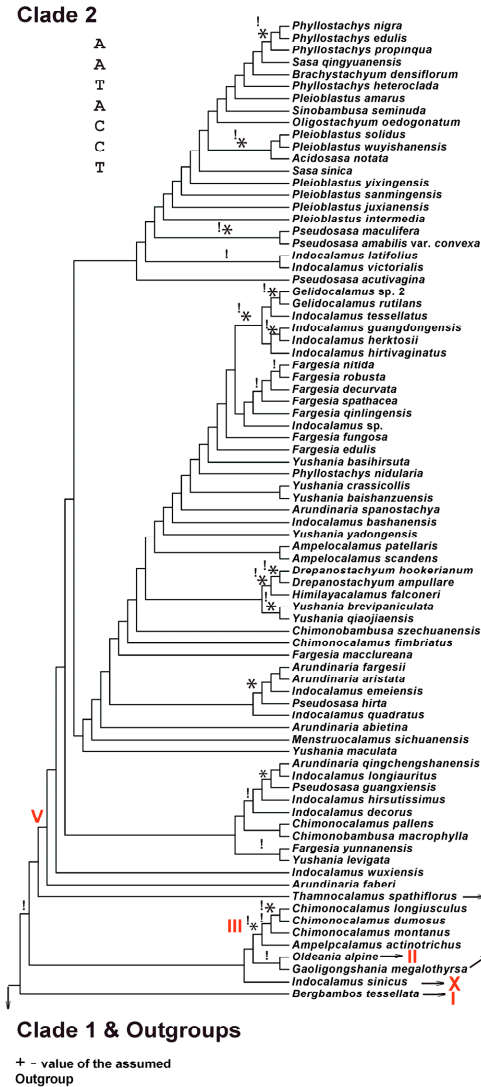
Given what is written above, we have developed *1001*, a computer program that converts unpolarized molecular data matrices into different types of binary matrices, either unpolarized or with established polarity.

1001 is implemented as a Python-based script that translates conventional matrices of unordered multisite characters into polarized and non-polarized binary matrices written in Phylip or Comma Separated Values (CSV) file formats. *1001* can be used with any operating system that has a Python interpreter (e.g., Linux, Mac OS X, and Windows) (<http://www.python.org/> accessed on 21 January 2025). Upon request, all basic formal algorithms for the binary representations of multistate characters originally designed by the first author were implemented in the Python-based script by Dr. Matthew A. Gitzendanner (University of Florida, Gainesville, FL 32611 USA).

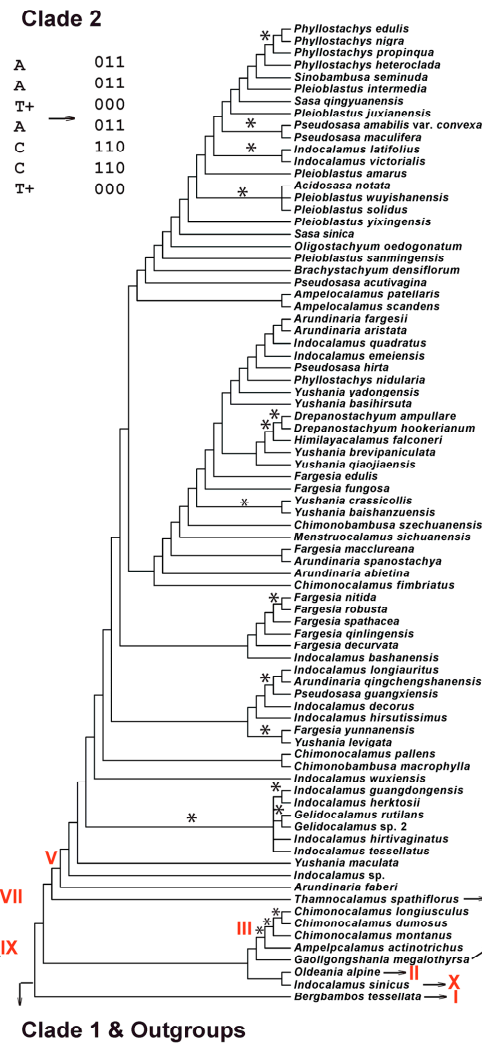
1001 accepts conventional DNA alignments in "relaxed" Phylip format [19,20]. All gaps and ambiguities of the conventional multistate matrices must be recoded as "?" ("missing entities") before running *1001*. The DNA sequence data are of our primary interest, and the default state of *1001* is designed for this kind of data. However, the script may handle different types of unordered multistate characters (amino acid sequence alignments and non-molecular data).

The implementation of *1001* ways of re-coding the DNA sequence data led to "non-direct methods of polarity estimation" [14,21–24]. Both methods implemented by *1001* (see below) *a priori* polarize conventional data by comparing them with an assumed all-plesiomorphic outgroup. Such cladistic methodology has been defined as an "out-group comparison" [4,12,14,25–27]. The detailed legends of Figures 1–4, as well as the four figures themselves, provide the summary and explanation of the results and methods of this study.

A. DNA characters



B. Binary data: Method 1
(polarized binary example)



C. Binary data: Method 2

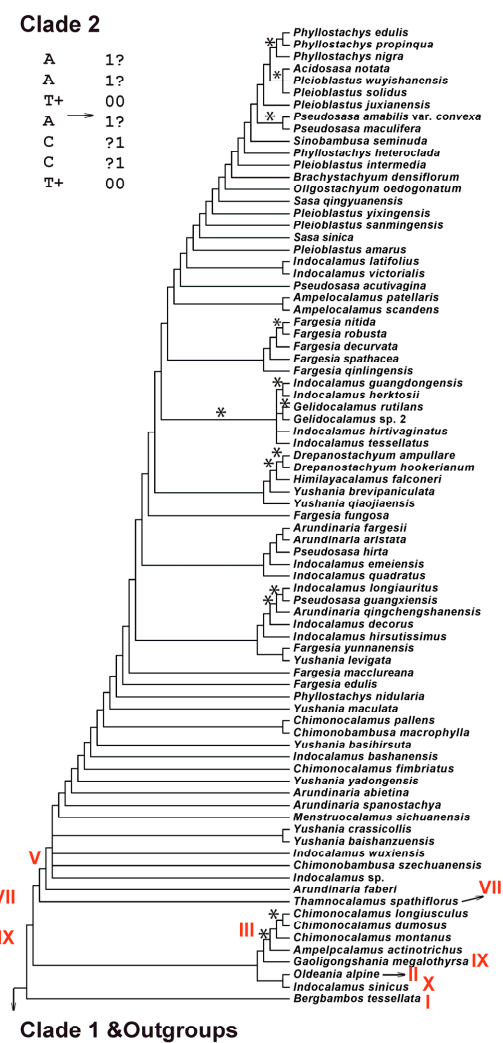
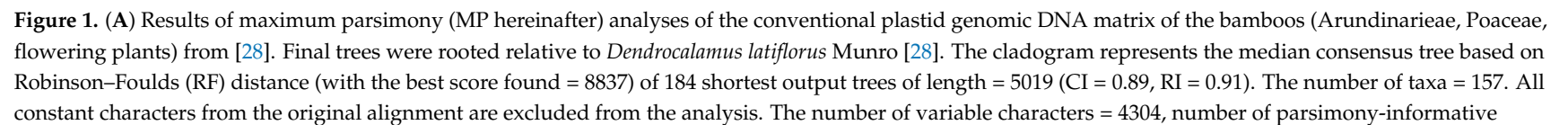


Figure 1. Cont.



characters = 2003. * nodes received MP Jackknife (JK) support > 50% after 20,000 fast JK replicates; ! nodes recovered MP Bootstrap support in the analysis from [28] (200 full heuristic replicates). **(B)** Results of MP analysis of the binary representation of the conventional DNA matrix from A., re-coded following the proposed 1001 Method 1. Initial binary data were polarized before analysis relative to *D. latiflorus*, assumed as an outgroup [28]. The cladogram represents the majority-rule consensus of 191 shortest output trees of length = 10,014 (CI = 0.88, RI = 0.89). The number of taxa = 157. The number of binary characters = 8783, number of parsimony-informative characters = 4088. * nodes received MP Jackknife (JK) support > 50% after 20,000 fast JK replicates. **(C)** Results of MP analysis of the binary representation of the conventional DNA matrix from A., re-coded following the proposed 1001 Method 2. Data polarized before analysis relative to *D. latiflorus*, assumed as an out-group based on the previous results of [28]. The cladogram represents the majority-rule consensus of 139 shortest output trees of length = 4993 (CI = 0.89, RI = 0.91). The number of taxa = 157. The number of binary characters = 4993, number of parsimony-informative characters = 2027. * nodes received MP Jackknife (JK) support > 50% after 20,000 fast JK replicates. All MP analyses were conducted using program PAUPrat [29–31] as implemented in CIPRES [32] following 200 ratchet replicates with no more than 10 trees of length greater than or equal to 1 saved in each replicate, and the TBR branch swapping/MulTrees option in effect; -pct = 20%, all characters weighted uniformly, and gaps were treated as “missing”. MP jackknifing [33] was conducted using PAUP* version 4.a168 [31] (PAUP* hereinafter) as implemented in CIPRES [32]. Robinson–Foulds consensus [14,34] calculated using RFS version 2.0 [34]. Majority-rule consensus calculated in PAUP* [31]. Branches with a minimum length of zero collapsed. All gaps and ambiguities of the conventional DNA matrix (A) were recoded as missing data (“?”) before binary permutations. Roman numerals correspond to the “major lineages” of Arundinarieae [28].

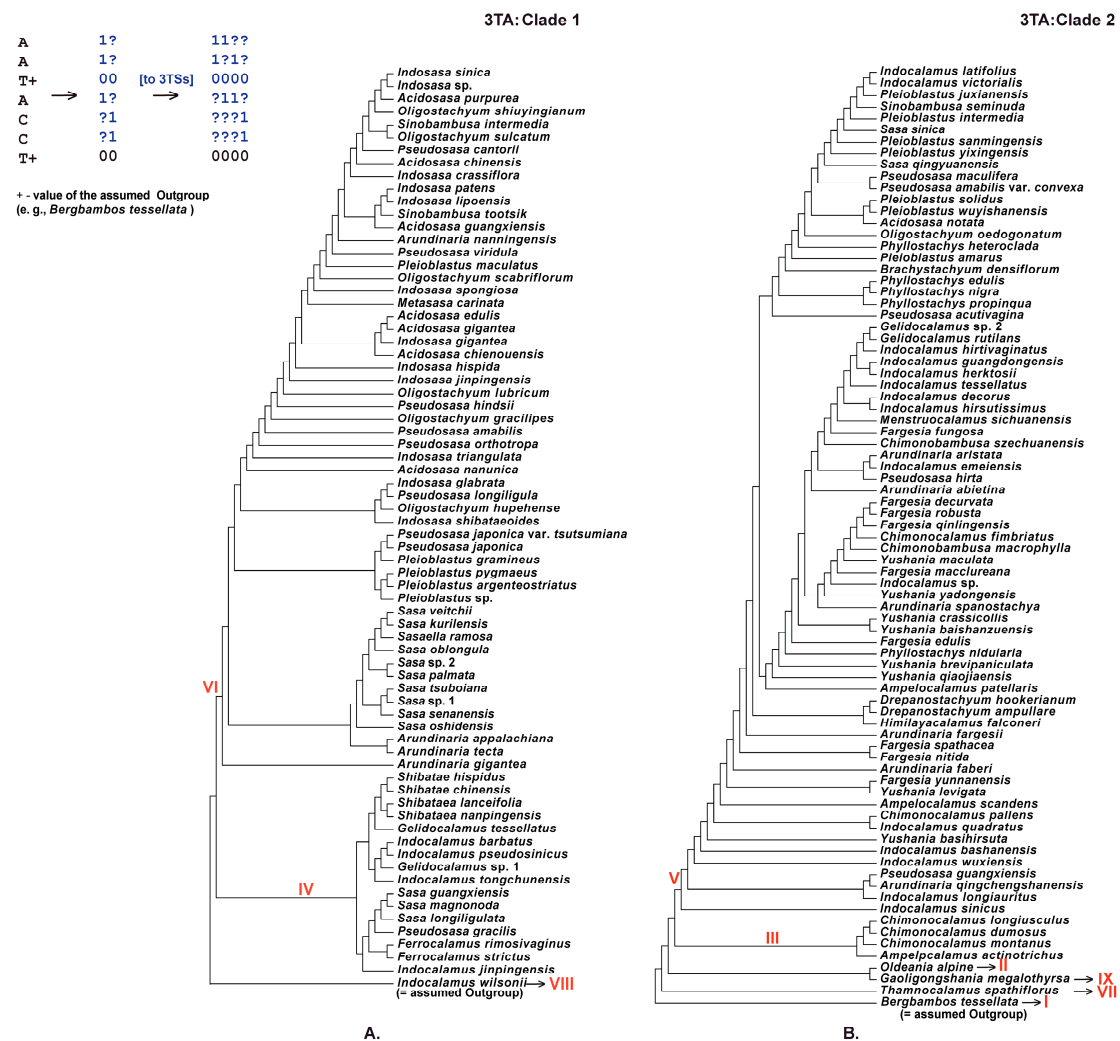


Figure 2. The results of two three-taxon statement analyses (3TA hereinafter) of Clades 1 and 2 (Figure 1). The DNA alignments have been polarized following 1001 Method 2 and subsequently established as binary three-taxon matrices using TAXODIUM version 1.2 [18] (TAXODIUM hereinafter). Following the results of the previous analyses (Figure 1), *Indocalamus wilsonii* (Rendle) C.S.Chao and C.D.Chu (Clade 1) and *Bergbambos tessellata* (Nees) Stapleton (Clade 2) were assumed to be outgroup taxa before Method 2 was applied to the DNA characters. **(A)** The results of the first 3TA (Clade 1). Majority-rule consensus of 193 shortest output trees of length = 527,046 (CI = 0.92, RI = 0.91). The number of taxa in the 487168 character–3TA matrix is 72. All 487,168 3TSs are parsimony-informative and weighted uniformly. **(B)** The results of the second 3TA (Clade 2). Majority-rule consensus of 201 shortest output trees of length = 187,857 (CI = 0.86, RI = 0.83). The number of taxa in the 161,027 character–3TA matrix is 80. All 1,610,278 3TSs are parsimony-informative and weighted uniformly. For the meaning of Roman numerals and the details of the MP analyses, see the legend of Figure 1.

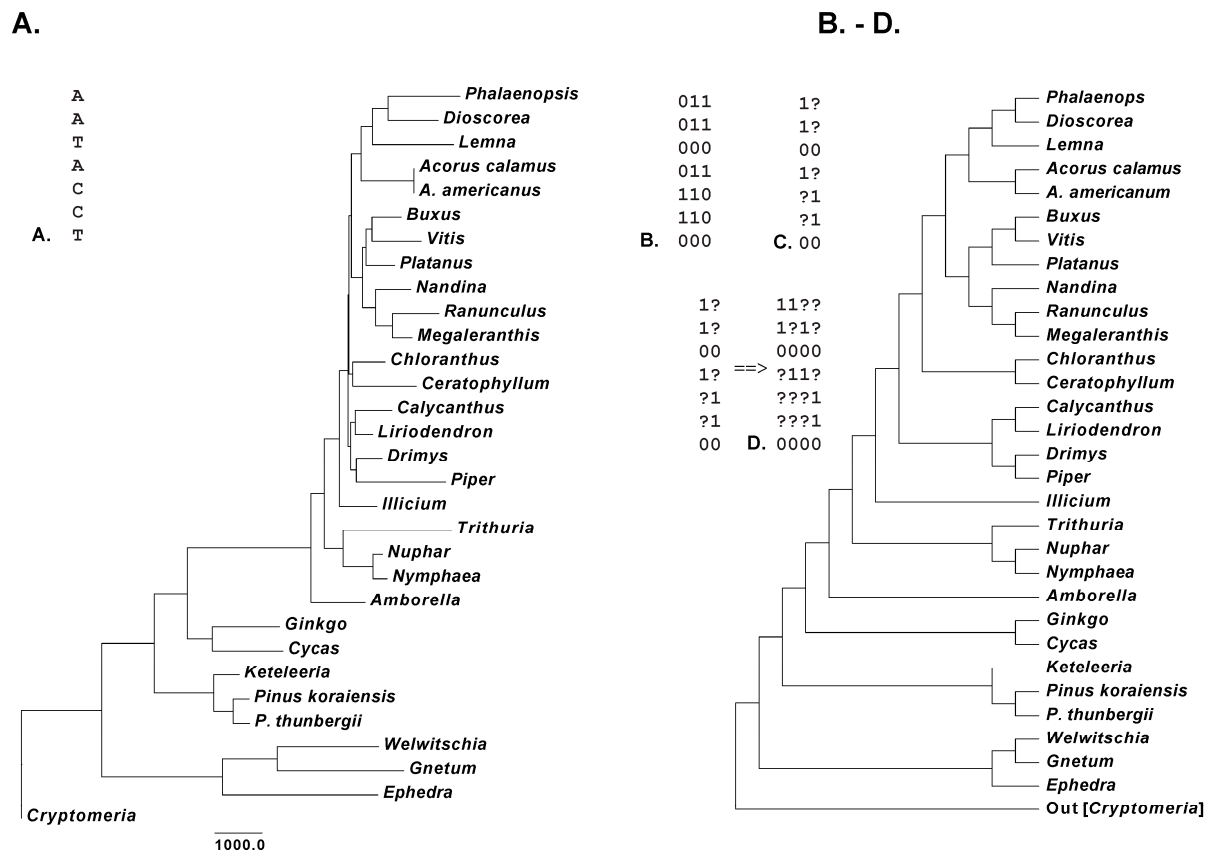


Figure 3. (A) The simplified phylogeny of flowering plants and outgroups resulted from the MP analysis of the 38,553 bp cpDNA alignment from [35]. The general strategy of the analysis is described in [18]. The heuristic search for the most parsimonious tree was performed with the implied weights [36] included in the search procedure, and the value of the k -function was assigned as three. The phylogeny is established as a single phylogram. Goloboff fit = $-10,023.39940$, with the actual length of the tree equal to 48186, CI = 0.55, RI = 0.61. The number of informative characters is equal to 13,328. (B) The most parsimonious hierarchy of patterns was obtained from the MP analysis of the same strategy as in (A). The latter was based on the polarized binary matrix recoded from the conventional cpDNA alignment (A) following 1001 Method 1, with *Cryptomeria* (Cupressaceae Bartlett, gymnosperms) assumed as the best outgroup. The hierarchy of patterns is established as a single cladogram. Goloboff fit = -24165.80162 with the actual length of the tree equal to 102,724, CI = 0.49, RI = 0.60. The number of informative characters equals 32,141. (C). The most parsimonious hierarchy of patterns resulted from the MP analysis, which followed the same strategy as in A (see above) but without implied weights [36] included in the search procedure. The analysis was based on the polarized binary matrix recoded from the conventional cpDNA alignment (A) following 1001 Method 2, assuming *Cryptomeria* as the best outgroup. The hierarchy of patterns is established as a single cladogram of the length 48,552, CI = 0.56, RI = 0.62. The number of informative characters equals 15,653. (D) The single most parsimonious hierarchy of patterns resulted from the MP analysis, which followed the same strategy as in (A) (see above) but without implied weights [36] included in the search procedure. The analysis was based on the three-taxon statement matrix with 1,652,888 fractionally weighted [4,12,14,18] three-taxon statements calculated by TAXODIUM [18]. This matrix is derived from the polarized binary representation (1001 Method 2) of the 28,196 bp largest clique, estimated by PHYLIP version 3.695 [19] based on a 38,553 bp cpDNA alignment (A). *Cryptomeria* is assumed to be the best outgroup. The hierarchy of patterns is established as a cladogram of the length of 230,181.7318, CI = 0.99, RI = 0.99. The number of informative characters (three-taxon statements) equals 1 652 888.

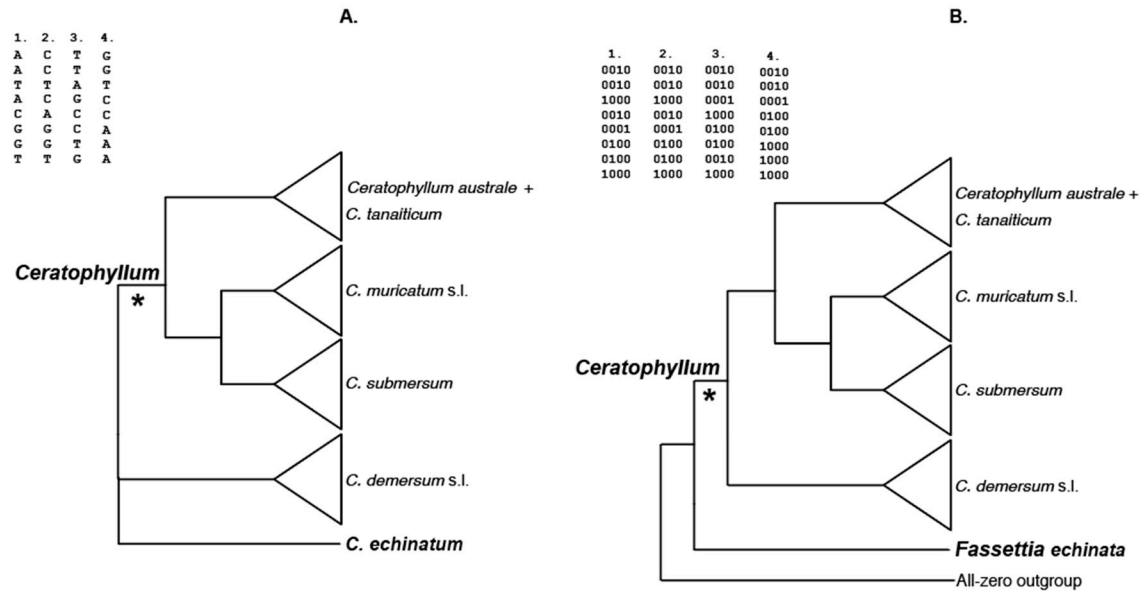


Figure 4. (A) An unrooted simplified molecular phylogeny of *Ceratophyllum* (Ceratophyllaceae A. Gray, flowering plants) [37], showing the ambiguous placement of *C. echinatum* [37]. (B) A summary of the cladistic analyses [38], demonstrating that *C. echinatum* is a sister group to the narrowly defined genus *Ceratophyllum*. All analyses (B) were based on the binary ‘presence–absence’ representation of the molecular data from [37], adding an artificial all-zero outgroup. As a result of the cladistic analyses of the binary recoded DNA sequence data [37,38], *C. echinatum* was defined as a sister group of the narrowly circumscribed genus *Ceratophyllum* [38] and transferred to the newly established genus *Fassettia* based on the obtained phylogenetic placement [38]. See [38] for details of the cladistic analyses and taxonomic treatment. Clade “*Ceratophyllum*” is marked with an asterisk (*). This figure also shows the ‘presence–absence’ binary coding (B) of the DNA sequence data (A), as implemented in 1001.

A few ways of binary coding of DNA data are possible [39,40]. The most straightforward one is frequently cited as the ‘Vos representation’ of DNA sequences (reviewed or implemented in [39–43], among others) or as “CODE-4 encoding” of DNA data [44]. The name of this technique in older cladistic literature is the “presence–absence” (or “absence–presence”) coding [14,45], and in this study we follow this latter naming.

The “presence–absence” binary encoding is intuitive and can be clarified using the following example of a binary recoded simple DNA character:

A = 0010

A = 0010

T = 0001

C = 1000

G = 0100

The latter example elucidates the implementation of ‘presence–absence’ coding in the context of previous versions of *1001*. However, encoding ‘A’ as 0001 and ‘T’ as 0010, among other possibilities, is entirely acceptable. These encoding choices should be examined independently. Due to the lack of such examination, in the current version of *1001*, the presence–absence coding procedure is applied not to the whole matrix but to each character individually. Consequently, within the entire output binary table, the “presence–absence” re-writing of the DNA alignment allows any nucleotide to be replaced by all four possible combinations of ones and zeros: 0001, 0010, 0100, and 1000 (Figure 4).

The “presence–absence” method assigns four symbols for each nucleotide, and all of these symbols are one and zero. Consequently, this method is more informative and accurate than other available techniques [39,40], which eventually attempt to encode each nucleotide using two symbols: 00, 01, 10, and 11. These methods are asymmetrical compared to the “presence–absence” framework: within the binary logic of two-digit DNA encoding, one nucleotide must be recoded exclusively as zeros, while the other must be recoded exclusively as ones.

For someone, the easiest way to polarize the binary matrix is to add the *artificial* all-zero outgroup to the “presence–absence” representation of the standard DNA-based alignment [14]. This general procedure may help resolve issues with unknown outgroups in specific cases [38]. However, it is critical to implement a more specific method that helps polarize the “presence–absence” binary matrix using binary recoded DNA sequence data from individuals of a real taxon (assumed outgroup). Both methods proposed in this paper address the latter issue Figures 1B and 3B).

Eight output files result from each run of *1001* if the first method is selected:

- (a) Non-polarized or ‘presence–absence’ binary matrix with no artificial all-zero outgroup added, with and without invariant characters (both Phylip (phy) and comma-separated values (CSV) files); The invariant (and non-informative) characters are, strictly speaking, not considered characters from the cladistic standpoint [2]. However, saving them in the resulting outputs may be necessary for future statistical analyses of the recoded DNA alignments. The same proposition is also valid for the second method of *1001* (see below). Using available software [13,20,31], we recommend removing all non-informative characters from the *1001* output files before conducting phylogenetic analyses.
- (b) Binary matrices that result from the polarization of the ‘presence–absence’ binary matrix relative to a real taxon (assumed outgroup), with and without invariant characters (both phy and CSV files).

It is also worth stressing that a single DNA sequence cannot be the subject of phylogenetic polarization and may be easily recoded to the ‘presence–absence’ binary format

using many available text editors, therefore, in principle, requiring no special software (see below).

The second method (Figure 1C and C), or as we prefer to call it, the “Cladistic” Method, directly represents the conventional DNA alignment of the parsimony-informative characters as a set of maximal relationships [2,8,27,46] following the values of the pre-selected outgroup taxon [2,8,27,46] (Figure 1C). The method is designed for the parsimony-informative inputs only.

Both proposed methods increase the number of parsimony-informative characters. The binary 1001 outputs can be used with popular phylogenetic software, with many different statistical packages, as well as an input for three-taxon permutations [47] using TAXODIUM (polarized binary data only) [18] for the future 3TA (Figures 2 and 3D). Such binary matrices can also be used as a source for rooted trees, generated by the FORESTER version 1.0 [27] and m2n version 1.0 [2] software, for future matrix-free cladistic explorations.

1001 is available for free from the web (<https://github.com/magitz/1001>) (accessed on 22 November 2024).

To show that even the straightforward binary ‘presence–absence’ recoding of DNA sequence data can be beneficial and enhance our understanding of the classification of specific groups of organisms, we provide a simple practical example (Figure 4). Due to the lack of a precisely known suitable outgroup for the flowering plant genus *Ceratophyllum* L. (Ceratophyllaceae A. Gray), Szalontai et al. (2018) [37] published an unrooted molecular phylogeny of the genus, which was reviewed in Mavrodiev et al. (2021) [38]. In a study by Szalontai et al. (2018) [37], the morphologically unique North American endemic species *C. echinatum* A. Gray was ambiguously placed (Figure 4A). The presence–absence binary recoding of the initial DNA sequence data [37], along with the future addition of a polarizing all-zero outgroup to the resulting binary table [38], demonstrated the sister relationship between *C. echinatum* and the other members of *Ceratophyllum* in a series of cladistic analyses [38]. Thus, using the presence–absence binary recoding of the DNA alignment, Mavrodiev et al. (2021) [38] predicted and confirmed that *C. echinatum* is a sister group to the narrowly defined *Ceratophyllum* (Figure 4B). As a result, *C. echinatum* was transferred to the newly established genus *Fassettia* Mavrodiev [38].

Many popular phylogenetic applications can polarize characters before analysis (e.g., the command “AncState” of PAUP* [13,31] and “ancstates” of TNT [48]; see also the option “ancestors” included in some programs of the PHYLIP package [19]. However, our analytical approach to DNA sequence data has never been properly investigated since the beginning of the molecular age [10,11]. If the characters of a conventional DNA alignment are polarized, then the data are represented in the form of *relations*, either “maximal” [2,4,8,27,46] or “minimal” [4,12,14,47]. One of the goals of 1001 is the explication of sets of maximal relationships as separate entities (as polarized binary matrices) for future analyses. Some semantic and analytical possibilities remain unexplored from this simple cladistic perspective of using polarized binary representations of conventional DNA alignments. Therefore, both proposed methods may enhance research within the field of molecular systematics. The listed popular software (see above) is unable to perform such explications, even if, in principle, these programs can polarize matrices before analyses.

By using unpolarized data, modern phylogeneticists are following Farris [4,22,49–55]. Meacham (1984) [54,55] also explicitly did not recommend polarizing characters before analysis. As was comprehensively summarized, it is better to infer the tree and the character polarities simultaneously, rather than going through the two-stage process of assigning polarities first and then estimating the tree [13,14,54–56]. However, due to this strategy, which requires an *a posteriori* rooting procedure, plesiomorphic characters are extensively optimized on the nodes of the inferred phylogenetic tree [4].

It was also noted that *a priori* polarization of the characters is reasonable only when the polarity determination is unambiguous (i.e., there is no heterogeneity in the outgroup for characters that are variable within the ingroup) [13,14,54–56]. When the outgroup is heterogeneous, the most parsimonious assignment of an ancestral condition for the ingroup depends upon how the outgroup taxa are related to each other [13,14,22,54–57]. In other words, if the number of taxa within the outgroup is in some way reduced to one [48,58], or in the case of a homogeneous outgroup, for a character with two or more states, the state occurring in the outgroup can indeed be assumed to be the plesiomorphic state [14,22–24,26,59,60].

After the binary representation of the DNA alignment, each column of the latter corresponds to one or a few columns of the new binary matrix (Figures 1–4). As we mentioned above, the semantics of these new columns of the binary table are different from those of the original DNA alignment. This is especially clear when the binary matrix is polarized, as already stated; in this case, the new binary characters represent the relationships of the taxa, whose individuals' DNA was sequenced. Again, such relations are no longer equal to the conventional DNA character but are the hypothesis, propositions, statements (either three or *n*-taxon statements), etc. [2,4,61,62]. Therefore, the polarized binary matrix is semantically different from either raw DNA-based alignment or the non-polarized binary representation of this alignment. One, for example, may note that the polarized binary matrix represents a kind of *structure* rather than a collection of raw characters.

The new denotation of the analyzed matrix suggests a revised framework for conducting phylogenetic analysis based on this matrix. For example, differences between the nucleotides in the conventional DNA alignment, if used as an input for phylogenetic software, imply some processes change nucleotides among each other (e.g., mutation from A to C). However, in the case of the polarized binary matrix derived from the conventional DNA-based alignment, a change from character state zero to character state one ($0 \Rightarrow 1$) does not necessarily associate with a nucleotide point base substitution. The meaning of this change is different.

As stated, a while ago,

'Application of absence/presence coding has yet to be considered in molecular systematics, and no body of opinion considers base substitution as anything other than a special form of character state transformation.' ([14], p. 36).

Today's mainstream phylogenetic view on nucleotide-based substitutions remains the same as articulated 27 years ago by Kitching et al. (1998) [14]. But as we see it now, phylogenetic analysis of polarized binary representations of the DNA alignments allows us to avoid base substitution (and the related idea of character state transformation) from an analytical perspective. The simple comparison of DNA sequences from individuals of different taxa is analytically sufficient if such sequences are established as inputs for phylogenetic analyses as polarized binary matrices. Therefore, the evolutionary interpretation of the tree results from the form of the input data representation. Thus, the objective of phylogenetic analysis of a polarized binary matrix derived from conventional DNA alignment is not to reconstruct historical events and evolutionary relationships among the analyzed taxa but to achieve their *hierarchical classification* [12].

Another may tell us that the notion that systematic data constitute a regular character $\times$ taxon matrix (e.g., the conventional DNA-based alignment) is not intrinsically cladistic [4,15,46,63]. Therefore, the different type of data may be required for cladistic analysis, especially if the latter is viewed as an extension of the comparative approach [4,8,12,64,65]. The three-taxon statement matrix is a relevant example of the cladistic data type [4,12,14,47,63]. The meaning of the latter matrix is a summary of all possible minimal (three-taxon) relations between analyzed taxa, based on the characters of the taxa's individuals [4,12,14,18,47]. The sets of the

other polarized binary representations of conventional DNA-based alignments (Figure 1B,C and Figure 3B,C) may also be considered good candidates for the proper inputs for cladistic analysis, at least within a matrix-based perspective.

One might argue that the methods of binary representation of DNA discussed in this paper are neither novel nor offer substantial benefits to the reader. For example, in R and MATLAB, there are built-in functions that can be slightly modified to easily implement the binary mapping of molecular data. This assertion is incorrect, not only because neither R nor MATLAB, as well as widely cited packages such as iLearn [66], iFeature [42], and SRAMP [41], have yet to implement the polarized binary recoding procedure (see above), regardless of the ability of the listed software to perform the ‘presence–absence’ recoding of the DNA structure. Therefore, both discussed methods implemented in 1001 are novel and beneficial for users. One such benefit is breaking with the long-standing tradition of using unpolarized conventional DNA-based alignments in molecular systematics.

Finally, it may be noted that following the implementation of a tree-based representation of conventional DNA alignment and the demonstration of the efficiency of matrix-free phylogenetic analyses [2,27,67], any future matrix-based binary representations of DNA data are deemed unnecessary. Indeed, after the DNA-based alignment is established as a forest of rooted trees, the latter may be easily converted to a binary matrix following the logic of the ‘Matrix Representation with Parsimony’ (MRP) approach [68,69] or other related methods. Different software is available to complete this task (e.g., [20]). Thus, the rooted tree-based representation of conventional DNA alignment [2,27,67] always implies a polarized binary matrix that can or cannot be saved as a separate entity.

We would respond that the MRP representation of standard DNA alignment differs from the most straightforward “presence–absence” recording of conventional DNA alignment. Such recording can benefit both cladistic analytics (Figure 4) and other perspectives related to the binary translation of the DNA alphabet. This is also true regarding the other binary methods discussed in this paper, but below, we will focus on the most unadorned ‘presence–absence’ DNA recording representation and demonstrate its advantages and relevance..

3. Binary Representation of the DNA Sequence Data, Leibniz, and Religion

In this essay, we aim to demonstrate that binary representations of the DNA molecules, in their simplest form of ‘presence–absence’ coding, can also broaden the understanding of DNA sequences themselves—even for the phylogenetic analyst, for whom DNA is merely a set of combinations of the four well-known basic symbols. The binary representations of DNA data have broad mathematical and cultural connotations and are intriguing, regardless of their applications to different phylogenetic procedures.

The key observation is that the “presence–absence” recoding, if applied to four aligned nucleotides, immediately links the standard four DNA symbols with Leibniz’s binary depiction (see below and Figure 5) of the four Arabic numbers, namely one, two, four, and eight, as follows:

A = 0001 = 1

T = 0010 = 2

C = 0100 = 4

G = 1000 = 8

Strictly speaking, the order of the nucleotides is unnecessary, and any of them can be recorded with any four-digit combination of the zeros and ones (see above). Noting that all four Arabic numerals belong to the geometric sequence 2^n , we can rewrite the latter representation in a more general form as follows:

$$N_1 = 0001 = 2^0$$

$$N_2 = 0010 = 2^1$$

$$N_3 = 0100 = 2^2$$

$$N_4 = 1000 = 2^3$$

where any **N** can be either **A**, **T**, **C**, or **G**.

The possibility of four-digit binary rewriting of the numbers one, two, three, and four is evident from Leibniz's texts, first, from his "*Explanation of Binary Arithmetic*" (1703) [70–72], as well as from his other numerous essays and letters, where such rewriting was established literally (Figure 5).

descendens in sinistro latere continet characteres a 0 ad 31, sed alter semicirculus rursus a summo descendens in sinistro latere continet characteres a 32 ad 63, ita enim sibi opponuntur e regione in eadem altitudine ii characteres qui non differunt nisi quod sinistrarum prima nota est 0 seu !, dextrarum vero 1 seu |. Ex. gratia 0 et 32, 1 et 33, 2 et 34

0	000000	100000	32
1	100000	100001	33
2	000010	100010	34
	etc.	etc.	

Ita mirum accidit, ut res ante ter et amplius annos nota in extremo nostri continentis oriente, nunc in extremo ejus occidente. sed melioribus ut spero auspiciis resuscitaretur. Nam non apparet, antea usum hujus characteresmi ad augendam numerorum scientiam innotuisse. Sinenses vero ipsi ne Arithmeticam quidem rationem intelligentes nescio quos mysticos significatus in chracteribus mere numeralibus sibi fingeant.*

* Leibniz hat bemerkt: Caetera alias prosequar, ostendamque generaliter summas seriei periodicae dare seriem periodicam, et columnis alttrius columnae summatrices habere periodum, hoc modo, ut ad minimum periodus primae columnae summatricis sit longitudine dupla periodi summandae, secundae longitudine quadrupla, tertiae longitudine octupla etc. Tunc enim semper 0 incidit in columnae summatricis et summatricium praecedentium simul cum fine periodi, ita omnia ab integro prodeunt ut ante. Hinc apparet statim periodus naturalium. Ostendam et addendo in unum duas columnas periodicas fieri periodicam. Ergo addendo in unum quocunque periodicas orietur periodica. Hinc demonstro numeros naturales et progressionis arithmeticae, et polygonos quosque, tum etiam quadraticos, cubicos etc. periodicas series dare.

0	0	--
1	1	—
00	0	==
01	1	—
10	2	==
11	3	—
000	0	===
001	1	==
010	2	==
011	3	==
100	4	==
101	5	==
110	6	==
111	7	==
0000	0	====
0001	1	==== !
0010	2	==== !
0011	3	==== !
0100	4	==== !
0101	5	====
0110	6	====
0111	7	==== !
1000	8	==== !
1001	9	====
1010	10	====
1011	11	====
1100	12	====
1101	13	====
1110	14	====
1111	15	====

Figure 5. Leibniz's original four-digit binary representation of Arabic numbers one, two, four, and eight (indicated by exclamation marks, added by us). In the third column of this table, Leibniz himself linked this representation with the combination of solid and dotted lines, each corresponding to one of the four *T'ai Hsüan Ching* tetragrams (indicated by exclamation marks, added by us), namely the tetragrams *Penetration*, *Legion*, *Fullness*, and *Law (Model)* [73]. Reproduced from Leibniz's manuscript *De Dyadics*, as interpreted and translated by Yakovlev [72], see pp. 195, 201, and 202.

From that perspective, any DNA chain can be represented by the combinations of natural numbers one, two, four, and eight, or by the first four members of the geometric progression 2^n , where n changes from zero to three, or by the above sequences of quadruples (tetrads) of ones and zeros. Thus, if connected with the Leibniz binary number system, the “presence–absence” coding of the DNA nucleotides implies that any DNA molecule in its sequence is not solely a word (an arrangement of letters) but an exact *number* that may be written in at least three different ways. Therefore, the most fundamental property of the DNA structure (the sequence of nucleotides) can be directly linked with either theoretical or applied aspects of arithmetic, algebra, or number theory.

Within the DNA alignment, some characters may consist of only three or two nucleotides. In these limited scenarios, the latter can be represented by binary triplets or pairs: 001, 010, and 100, or 01 and 10. The corresponding Arabic numerals for these representations are one, two, and four or one and two, respectively (Figure 5). This circumstance does not negate the general method of describing the DNA molecule using quadruples of zeros and ones, along with the corresponding Arabic numerals. Reducing the number of zeros and ones from four to three allows the technical avoidance of additional all-zero columns in the output binary table.

One may note that binary representations of DNA are a well-established practice today [74,75], particularly in visualizing macromolecules [75]. Consequently, the examples provided in this article are trivial. We would argue that the proper connection of this study lies not in the modern applied aspects of computer science and related disciplines (though they are essential), but in Leibniz’s original vision of the binary concept (see below). This symbolic vision clearly illustrates the highly productive synthetic scientific perspective that has nearly vanished from the modern high-tech scientific landscape.

The publication history on the topic of ‘Leibniz and binary’ is extensive [70–72,76], As summarized by Strickland (2024) ([70], p. 1):

[Leibniz] “... enduring mathematical contributions was his invention of the binary number system the basis for today’s world of digital computing and communications. In this system, which uses only two digits—0 and 1—every position starting from the right represents a successive power of 2...”

The same author stressed,

“Leibniz’s faith in the illuminatory power of binary is nowhere more evident than in his use of it in philosophical theology. In 1694/5, when sketching notes on one of Weigel’s books, Leibniz devised an analogy between the representation of all numbers by 1 and 0 and the theologically orthodox idea of creation ex nihilo, that is, the creation of all things out of nothing by God. Treating God as the analogue of 1 and nothingness as the analogue of 0, Leibniz took the origination of all numbers from 1 and 0 in binary as a reflection of the doctrine that all created things originate from God and nothingness. This analogy would play a pivotal role in Leibniz’s willingness to inform correspondents about binary from 1696 onwards (and ... played a pivotal role in his decision to publish an essay on binary in 1705)”. ([70], p. 1).

In short, for Leibniz himself, the binary number system had deep philosophical and religious connotations, but it is also worth stressing that the mathematical ‘... basis for today’s world of digital computing and communications’ was inspired in Leibniz’s mind by ‘the theologically orthodox idea of creation ex nihilo, that is, the creation of all things out of nothing by God’ [70]. Thus, Leibniz’s writings on binary numbers can be viewed as an example of the scientific heuristic value of the religious context of his research.

Wolfgang Pauli [77] described a similar case in his essay on Johannes Kepler (1571–1630), who was almost Leibniz’s older coeval, an outstanding mathematician and

one of the most prominent astronomers in the history of science. Among other things, in this text, the co-founder of modern quantum physics linked Kepler's strong heliocentric views with Kepler's symbolism of the Holy Trinity [77].

Returning to the main topic of this paper, we may suggest that Leibniz's views on binary represent a connection between the binary-encoded DNA structure and the world of religious symbols and doctrines. The latter connection is multivalent and is not limited solely to Leibniz's reference to the biblical notion of creation *ex nihilo*, the prototype of his binary universe. Following the comprehensive opinion of Joachim Bouvet (1656–1730), Jesuit missionary in Beijing, China, Leibniz himself completely equated his binary number system with the basic linear symbols found in a cornerstone text for Confucian and Daoist philosophical movements, *The Book of Changes* (*Yijing* or *I Ching*) (*Yijing* hereinafter), as conceptualized by the Chinese philosopher Shao Yong in the 11th century [70–72,76,78]. The connection between linear symbols of *Yijing* and Leibniz's binary notation, as established by Leibniz himself, is reproduced in Figure 5. As it perhaps clears from this figure, the solid line means one, and the dotted line means zero. The combinations of solid and dotted lines are equal to the combinations of one and zero or *yin* and *yang*, using the traditional language of Daoist philosophy [78].

In the minds of both Bouvet and Leibniz, the biblical notion of creation *ex nihilo* elucidates the 'binary' tradition of the *Yijing* and the other texts of the High Chinese Classics [70–72,78]. However, Leibniz's perspective on the semantic symmetry between his binary numbering system and the symbolic language of the *Yijing* was rational, primarily because he embraced the biblical concept of the intrinsic value of numbers in the created world. Leibniz explicitly pointed out that ancient Chinese authors did indeed discover the binary system of numbers and mistakenly associated it with their unnecessary religious mysticism [70–72,78].

Unfortunately, the Leibnizian rational dimension sometimes becomes blurred in some scientific, semi-scientific, and popular studies of differing quality that attempt to establish exact connections between the 64 triplets of the genetic code and the 64 hexagrams of the *Yijing*. This body of research still awaits evaluation. The 'presence-absence' binary expressions of the four nucleotides suggest that such attempts may be misguided.

To illustrate this potential misguidance, it is critical to repeat that the representations of the numbers one, two, four, and eight are four-digit sequences (Figure 5). Consequently, the *tetragrams*—comprising the tetrads of solid and dotted lines from the symbolic language of *Yijing*—should correspond to the four DNA nucleotides. However, the *Yijing* is based on eight trigrams (the triads of solid and dotted lines) and 64 hexagrams (the hexads of solid and dotted lines) [70–72,78]. To align with the hexagram-based framework of *Yijing*, DNA nucleotides should be encoded using two symbols (11, 10, 01, and 00) rather than four, as in the presence-absence method. Only when using two symbols to encode each nucleotide, the triplet of nucleotides is equivalent to six binary digits or to a combination of six lines, either solid or dotted, as in *Yijing*. But as previously mentioned, this coding approach has disadvantages compared to the four-symbol-based "presence-absence" method: it is asymmetrical and less informative. Thus, it is better to represent the nucleotides using quadruples of ones and zeros.

The latter observation suggests that the four-digit "presence-absence" binary representations of the nucleotides should be linked not with the *Yijing*, but with another text from the same philosophical tradition, the text of *T'ai hsuan ching* [73], which uses tetragrams rather than hexagrams, as in the *Yijing* [73], as its symbolic language. From this, the entire idea of connecting the 64 hexagrams of the *Yijing* with the 64 nucleotide triplets seems problematic. Such a connection may be technically possible through the correspondence between the tetragrams of the *T'ai hsuan ching* and the hexagrams of the

Yijing, as established in the texts of the *T'ai hsuan ching* [73]. However, this explanation has yet to be rationally approached.

Although Leibniz did not explicitly mention the *T'ai Hsüan ching* in his writings on binary systems (likely due to the non-binary nature of the general symbolism of this literary monument [73]), we found that his texts do literally include, among others, four specific *binary* tetragrams from the *T'ai Hsüan ching*—namely, *Penetration*, *Legion*, *Fullness*, and *Law (Model)* (see [73] for the composition and review of these symbols of Daoist tradition) (Figure 5). These tetragrams could, in principle, be associated with the four nucleotides recoded as 0001, 0010, 0100, and 1000 (Figure 5) following the “presence–absence” coding procedure. However, as mentioned above, it is impossible to determine which sequence of four zeros and ones (and thus which *T'ai hsuan ching* tetragram) corresponds to each nucleotide precisely. Any assumed exact correspondence is speculative. Therefore, this discussion focuses solely on the *principal* coincidence between four *T'ai hsuan ching* tetragrams, binary quadruples, and the four-letter alphabet of DNA, rather than on exact solutions, which are likely unavailable. Consequently, establishing any precise symmetry between the triplets of *T'ai hsuan ching* tetragrams and the triplets of nucleotides in the genetic code is also impossible.

4. Discussion

Discussing forgotten and lost perspectives is always a turn towards the future. As illustrated earlier, the binary representation of DNA sequence alignments may extend the horizon of conventional molecular phylogenetics, even changing its semantics. Within the binary world, character–state transformation, as given within the analytical algorithms, such as determining the most parsimonious sequence of character changes on a tree, is not necessarily related to natural processes, such as nucleotide substitutions. Therefore, binary representations of conventional DNA alignments enable investigators to analyze molecular characters from a purely comparative, even static, perspective. To reiterate, such representations simultaneously sharpen the research focus of the phylogenetic study on the classification dimension of the resulting cladogram, which is not necessarily related to the unknown evolutionary history of the studied taxa.

One might argue that citing numerous old references, as we did above, is pointless and anachronistic. Discussing and citing old literature, we remind the reader that many topics related to the foundations of phylogenetic analysis are yet to be resolved [4,12]. The almost complete dissolving of the binary perspective from modern molecular phylogenetics is a part of the same phenomenon. Despite the prevailing scientific consensus on the dominance of large-scale genomic data and parametric methods, such as the maximum likelihood approach [6,79] in modern industrial molecular phylogenetics, this consensus represents only one of the potential historical pathways explored in this scientific branch over the past decades. The observations discussed above pertain not only to the history and sociology of modern biological science but also to the future of phylogenetic analyses. Accurate binary DNA encoding of conventional DNA alignment will eventually enhance the accuracy of these analyses and prompt a rethinking of their underlying philosophy. Specifically, such encoding clarifies that the true focus of phylogenetic analyses is not the data themselves but the various representations of them. The latter also serves as a reminder of the comparative or static context in which DNA sequence data phylogenetic analyses can be conducted.

To retreat, in this study the ‘presence–absence’ coding procedure is applied not to the whole molecular matrix but to each character of the DNA alignment *individually*. Consequently, within the entire output binary table, the ‘presence–absence’ recoding of the DNA alignment allows any nucleotide to be replaced by all four possible combinations of ones and zeros: 0001, 0010, 0100, and 1000. For example, nucleotide A is recorded by all

four listed four-digit combinations, as shown in Figure 4B. From that, there are no ‘absolute’ links between, for example, the symbol ‘A’ (adenine) and the digital sequence 0001, and therefore between the same symbol ‘A’ and the Arabic numeral one. This observation suggests that Arabic numerals represent a complimentary form of binary notation (Figure 5). Consequently, these circumstances alone suffice to assert that the correspondence between Arabic numerals and DNA nucleotides is unlikely to be interpreted as biological. A similar conclusion applies to the connection between the *T'ai Hsuan ching* tetragrams and the standard symbols of DNA nucleotides, which becomes evident following the binary recoding of DNA sequence data (Figure 5). It is important to note that the biological functions of DNA, as well as other linear molecules such as RNA or proteins, are determined by their chemical and three-dimensional structures, rather than by any combination of numbers or symbols. Therefore, the observed coincidences among sequences of zeros and ones, Arabic numerals, nucleotides, and cultural symbols and concepts (Figure 5) are primarily of formal, esthetic, computational, and other significance, rather than strictly biological.

The straightforward interpretation of the description of DNA sequences using four natural numbers may take some time to be understood even from a computational standpoint. However, such a ‘delay’ is not a reason to ignore this possibility. For example, in 1982, Felsenstein and co-authors described a formal method for computing the fraction of matches between two nucleic acid sequences at all possible alignments [43]. This method employs the Fourier transform (reviewed in [43,80]). It also requires recording RNA molecules in digital form following the modified “presence–absence” coding procedure [43] (p. 134), thereby implying the Leibnizian binary context. Despite the title of their paper, “An Efficient Method for Matching Nucleic Acid Sequences”, the paper contains no straightforward biological examples [43], and the relevance of their approach may be easily questioned. Only 20 years after the publication of this text [43], the multiple sequence alignment program MAFFT, based on the Fourier transform, was established [80]. The development of MAFFT [80], now one of the most efficient multiple alignment software packages, also took over a decade of work [81].

In final paragraph, we would like to emphasize again, that the cultural and philosophical implications of the convergence between scientific and symbolic languages in describing the sequence of the DNA molecule should not be underestimated. The correlation of DNA sequence research with a Jungian archetypal perspective [77,82], as derived from the observations above, represents merely one of many contexts in which the intersection between the language of DNA and religious symbolism occurs.

Author Contributions: Conceptualization, E.V.M.; methodology, E.V.M.; validation, E.V.M. and N.E.M.; formal analysis, E.V.M.; investigation, E.V.M. and N.E.M.; resources, E.V.M.; writing—original draft preparation, E.V.M. and N.E.M.; writing—review and editing, E.V.M. and N.E.M.; supervision, E.V.M.; funding acquisition, E.V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Matthew Gitzendanner (University of Florida) for his elegant Python-based implementation of 1001. We thank both anonymous reviewers for their valuable comments. No agreement with any suggestions, statements, or conclusions of this paper is implied on behalf of any person mentioned or implied in this section. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zuntini, A.R.; Carruthers, T.; Maurin, O.; Bailey, P.C.; Leempoel, K.; Brewer, G.E.; Epitawalage, N.; Françoso, E.; Gallego-Paramo, B.; McGinnie, C.; et al. Phylogenomics and the rise of the Angiosperms. *Nature* **2024**, *629*, 843–850. [CrossRef] [PubMed]
2. Mavrodiev, E.V.; Madorsky, A. On pattern–cladistic analyses based on complete plastid genome sequences. *Acta Biotheor.* **2023**, *71*, 22. [CrossRef] [PubMed]
3. Hennig, W. *Phylogenetic Systematics*; University of Illinois Press: Urbana, IL, USA, 1966.
4. Williams, D.M.; Ebach, M.C. *Foundations of Systematics and Biogeography*; Springer: New York, NY, USA, 2008.
5. Farris, J.S. The logical basis of phylogenetic analysis. In *Advances in Cladistics*, 2. *Proceedings of the 2nd Meeting of the Willi Hennig Society*; Platnick, N.I., Funk, V., Eds.; Columbia University Press: New York, NY, USA, 1983; pp. 7–36.
6. Felsenstein, J. *Inferring Phylogenies*; Sinauer Associates Inc.: Sunderland, MA, USA, 2004.
7. Rannala, B.; Yang, Z.H. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* **1996**, *43*, 304–311. [CrossRef]
8. Nelson, G.J.; Platnick, N. *Systematics and Biogeography: Cladistics and Vicariance*; Columbia University Press: New York, NY, USA, 1981.
9. Platnick, N.I. Philosophy and the transformation of cladistics revisited. *Cladistics* **1985**, *1*, 87–94. [CrossRef] [PubMed]
10. Wägele, J.W. Hennig’s phylogenetic systematics brought up to date. In *Milestones in Systematics*; Williams, D.M., Forey, P.L., Eds.; CRC Press: Boca Raton, FL, USA, 2004; pp. 101–125.
11. Wägele, J.W. *Foundations of Phylogenetic Systematics*; Pfeil Verlag: München, Germany, 2005.
12. Williams, D.M.; Ebach, M.C. *Cladistics. A Guide to Biological Classification*, 3rd ed.; Systematics Association Special; Cambridge University Press: Cambridge, UK, 2020; Volume Series 88.
13. Swofford, D.L.; Begle, D.P. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 3.1. User’s Manual*; Laboratory of Molecular Systematics, MRC 534, MSC, Smithsonian Institution: Washington, DC, USA, 1993.
14. Kitching, I.J.; Forey, P.L.; Humphries, C.; Williams, D. *Cladistics: The Theory and Practice of Parsimony Analysis*; Oxford University Press: Oxford, UK, 1998.
15. Ebach, M.C.; Williams, D.M.; Vanderlaan, T.A. Implementation as theory, hierarchy as transformation, homology as synapomorphy. *Zootaxa* **2013**, *3641*, 587–594. [CrossRef]
16. Wiley, E.O.; Lieberman, B.S. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*, 2nd ed.; Wiley-Blackwell: Hoboken, NJ, USA, 2011.
17. Williams, D.M.; Siebert, D.J. Characters, homology and three–item analysis. In *Homology and Systematics: Coding Characters for Phylogenetic Analysis*; Scotland, R.W., Pennington, R.T., Eds.; Systematics Association Special Volumes; Chapman and Hall: London, UK; New York, NY, USA, 2000; Volume 58, pp. 183–208.
18. Mavrodiev, E.V.; Madorsky, A. TAXODIUM Version 1.0: A simple way to generate uniform and fractionally weighted three–item matrices from various kinds of biological data. *PLoS ONE* **2012**, *7*, e48813. [CrossRef]
19. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **1989**, *5*, 164–166.
20. Maddison, W.P.; Maddison, D.R. Mesquite: A Modular System for Evolutionary Analysis Version 3.81. Available online: <https://www.mesquiteproject.org/> (accessed on 12 October 2024).
21. Nelson, G.J. Ontogeny, phylogeny, paleontology, and Biogenetic Law. *Syst. Zool.* **1978**, *27*, 324–345. [CrossRef]
22. Nixon, K.C.; Carpenter, J.M. On outgroups. *Cladistics* **1993**, *9*, 413–426. [CrossRef]
23. de Pinna, M.C.C. Ontogeny, rooting, and polarity. In *Models in Phylogeny Reconstruction*; Scotland, R.W., Siebert, D.J., Williams, D.M., Eds.; Systematics Association Special Volume Series; Clarendon Press: Oxford, UK, 1994; Volume 52, pp. 157–172.
24. Bryant, H.N. Character polarity and the rooting of cladograms. In *The Character Concept in Evolutionary Biology*; Wagner, G.P., Ed.; Academic Press: San Diego, CA, USA, 2001; pp. 319–342.
25. Wiley, E.O. *The phylogeny and biogeography of fossil and recent gars (Actinopterygii: Lepisosteidae)*; Miscellaneous Publication—; University of Kansas, Museum of Natural History: Lawrence, KS, USA, 1976; Volume 64, pp. 1–111.
26. Platnick, N.I.; Gertsch, W.J. *The Suborders of Spiders: A Cladistic Analysis (Arachnida, Araneae)*; American Museum of Natural History: New York, NY, USA, 1976; Number 2607; pp. 1–15.
27. Mavrodiev, E.V.; Dell, C.; Schroder, L. A laid-back trip through the Hennigian Forests. *PeerJ* **2017**, *5*, e3578. [CrossRef]
28. Ma, P.-F.; Zhang, Y.-X.; Zeng, C.-X.; Guo, Z.-H.; Li, D.-Z. Chloroplast phylogenomic analyses resolve deep–level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). *Syst. Biol.* **2014**, *63*, 933–950. [CrossRef] [PubMed]
29. Nixon, K.C. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* **1999**, *15*, 407–414. [CrossRef] [PubMed]
30. Sikes, D.S.; Lewis, P.O. *PAUPRat: PAUP* Implementation of the Parsimony Ratchet. Beta Software, Version 1*; Distributed by the Authors; Department of Ecology and Evolutionary Biology, University of Connecticut: Storrs, CT, USA, 2001.

31. Swofford, D.L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.0b10*; Sinauer Associates: Sunderland, MA, USA, 2002.
32. Miller, M.A.; Pfeiffer, W.; Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, USA, 14 November 2010; Saltz, J., Ed.; IEEE: New Orleans, LA, USA, 2010.
33. Farris, J.S.; Albert, V.A.; Källersjö, M.; Lipscomb, D.; Kluge, A.G. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **1996**, *12*, 99–124.
34. Bansal, M.S.; Burleigh, J.G.; Eulenstein, O.; Fernandez-Baca, D. Robinson–Foulds supertrees. *Algorithms Mol. Biol.* **2010**, *5*, 18. [[CrossRef](#)]
35. Goremykin, V.V.; Nikiforova, S.V.; Biggs, P.J.; Zhong, B.; Delange, P.; Martin, W.; Woetzel, S.; Atherton, R.A.; Mclenachan, P.A.; Lockhart, P.J. The evolutionary root of flowering plants. *Syst. Biol.* **2013**, *62*, 50–61. [[CrossRef](#)]
36. Goloboff, P.A. Estimating character weights during tree search. *Cladistics* **1993**, *9*, 83–91. [[CrossRef](#)]
37. Szalontai, B.; Stranzinger, S.; Mesterhazy, A.; Scribailo, R.W.; Les, D.H.; Efremov, A.N.; Jacono, C.C.; Kipriyanova, L.M.; Kaushik, K.; Laktionov, A.P.; et al. Molecular phylogenetic analysis of *Ceratophyllum* L. taxa: A new perspective. *Bot. J. Linn. Soc.* **2018**, *188*, 161–172. [[CrossRef](#)]
38. Mavrodiev, E.V.; Williams, D.M.; Ebach, M.C.; Mavrodieva, A.E. *Fassettia*, a new North American genus of family Ceratophyllaceae: Evidence based on cladistic analyses of current molecular data of *Ceratophyllum*. *Aust. Syst. Bot.* **2021**, *34*, 431–437. [[CrossRef](#)]
39. Bernaola-Galvan, P.; Carpena, P.; Roman-Roldan, R.; Oliver, J.L. Study of statistical correlations in DNA sequences. *Gene* **2002**, *300*, 105–115. [[CrossRef](#)]
40. Mendizabal-Ruiz, G.; Román-Godínez, I.; Torres-Ramos, S.; Salido-Ruiz, R.A.; Morales, J.A. On DNA numerical representations for genomic similarity computation. *PLoS ONE* **2017**, *12*, e0173288. [[CrossRef](#)]
41. Zhou, Y.; Zeng, P.; Li, Y.H.; Zhang, Z.; Cui, Q. SRAMP: Prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* **2016**, *44*, e91. [[CrossRef](#)] [[PubMed](#)]
42. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.C.; et al. iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499–2502. [[CrossRef](#)] [[PubMed](#)]
43. Felsenstein, J.; Sawyer, S.; Kochin, R. An efficient method for matching nucleic acid sequences. *Nucleic Acids Res.* **1982**, *10*, 133–139. [[CrossRef](#)] [[PubMed](#)]
44. Demeler, B.; Zhou, G.W. Neural network optimization for *Escherichia coli* promoter prediction. *Nucleic Acids Res.* **1991**, *19*, 1593–1599. [[CrossRef](#)]
45. Pleijel, F. On character coding for phylogeny reconstruction. *Cladistics* **1995**, *11*, 309–315. [[CrossRef](#)]
46. Williams, D.M.; Ebach, M.C. The data matrix. *Geodiversitas* **2006**, *28*, 409–420.
47. Nelson, G.J.; Platnick, N.I. Three–taxon statements—A more precise use of parsimony? *Cladistics* **1991**, *7*, 351–366. [[CrossRef](#)]
48. Goloboff, P.A.; Farris, J.S.; Nixon, K.C. TNT, a free program for phylogenetic analysis. *Cladistics* **2008**, *24*, 774–786. [[CrossRef](#)]
49. Nixon, K.C.; Carpenter, J.M. On homology. *Cladistics* **2012**, *28*, 160–169. [[CrossRef](#)]
50. Farris, J.S. Methods for computing Wagner trees. *Syst. Zool.* **1970**, *19*, 83–92. [[CrossRef](#)]
51. Farris, J.S. Estimating phylogenetic trees from distance matrices. *Am. Nat.* **1972**, *106*, 645–668. [[CrossRef](#)]
52. Farris, J.S. Outgroups and parsimony. *Syst. Zool.* **1982**, *31*, 328–334. [[CrossRef](#)]
53. Kluge, A.G. Phylogenetic relationships in the lizard family Pygopodidae: An evaluation of theory, methods and data. *Miscellaneous Publ. Mus. Zool. Univ. Mich.* **1976**, *152*, 1–72.
54. Meacham, C.A. The role of hypothesized direction of characters in the estimation of evolutionary history. *Taxon* **1984**, *33*, 26–38. [[CrossRef](#)]
55. Meacham, C.A. Polarity assessment in phylogenetic systematics—More about directed characters—A reply. *Taxon* **1986**, *35*, 538–540. [[CrossRef](#)]
56. Maddison, W.P.; Donoghue, M.J.; Maddison, D.R. Outgroup analysis and parsimony. *Syst. Zool.* **1984**, *33*, 83–103. [[CrossRef](#)]
57. Lyons-Weiler, J.; Hoelzer, G.A.; Tausch, R.J. Optimal outgroup analysis. *Biol. J. Linn. Soc.* **1998**, *64*, 493–511. [[CrossRef](#)]
58. Arnold, E.N. Systematics and adaptive radiation of equatorial African lizards assigned to the genera *Adolfus*, *Bedriagaia*, *Gastropholis*, *Holaspis*, and *Lacerta* (Reptilia, Lacertidae). *J. Nat. Hist.* **1989**, *23*, 525–555. [[CrossRef](#)]
59. Watrous, L.E.; Wheeler, Q.D. The out-group comparison method of character analysis. *Syst. Zool.* **1981**, *30*, 1–11. [[CrossRef](#)]
60. Donoghue, M.J.; Maddison, W.P. Polarity assessment in phylogenetic systematics: A response to Meacham. *Taxon* **1986**, *35*, 534–538. [[CrossRef](#)]
61. Platnick, N.I.; Humphries, C.J.; Nelson, G.; Williams, D.M. Is Farris optimization perfect?: Three–taxon statements and multiple branching. *Cladistics* **1996**, *12*, 243–252. [[PubMed](#)]
62. Wilkinson, M. Common cladistic information and its consensus representation: Reduced Adams and reduced cladistic consensus trees and profiles. *Syst. Biol.* **1994**, *43*, 343–368. [[CrossRef](#)]

63. Platnick, N.I. Character optimization and weighting—Differences between the standard and three–taxon approaches to phylogenetic inference. *Cladistics* **1993**, *9*, 267–272. [[PubMed](#)]
64. Rieppel, O.; Williams, D.M.; Ebach, M.C. Adolf Naef (1883–1949): On foundational concepts and principles of systematic morphology. *J. Hist. Biol.* **2013**, *46*, 445–510. [[CrossRef](#)]
65. Nelson, G.J. Outline of a theory of comparative biology. *Syst. Zool.* **1970**, *19*, 373–384. [[CrossRef](#)]
66. Chen, Z.; Zhao, P.; Li, F.; Marquez-Lago, T.T.; Leier, A.; Revote, J.; Zhu, Y.; Powell, D.R.; Akutsu, T.; Webb, G.I.; et al. iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.* **2020**, *21*, 1047–1057. [[CrossRef](#)]
67. Mavrodiev, E.V.; Williams, D.M.; Ebach, M.C. On the typology of relations. *Evol. Biol.* **2019**, *46*, 71–89. [[CrossRef](#)]
68. Baum, B.R. Combining trees as a way of combining data sets for phylogenetic inference and the desirability of combining gene trees. *Taxon* **1992**, *41*, 3–10. [[CrossRef](#)]
69. Ragan, M.A. Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *BioSystems* **1992**, *28*, 47–55. [[CrossRef](#)]
70. Strickland, L. Leibniz on number systems. In *Handbook of the History and Philosophy of Mathematical Practice*; Sriraman, B., Ed.; Springer: Cham, Switzerland, 2024; pp. 167–197.
71. Strickland, L.; Lewis, H.R. *Leibniz on Binary: The Invention of Computer Arithmetic*; MIT Press: Boston, MA, USA, 2022.
72. Yakovlev, V.M. *Leibniz G.W.: Letters and Essays on Chinese Philosophy and the Binary System of Calculation (Preface, Translations, and Notes)*; Russian Academy of Sciences, Institute of Philosophy: Moscow, Russia, 2005.
73. Nylan, M. *The Canon of Supreme Mystery, by Yang Hsiung. A Translation with Commentary of the T'ai Hsuan Ching*; State University of New York Press: Albany, NY, USA, 1993.
74. Chen, W.; Liao, B.; Xiang, X.; Zhu, W. An improved binary representation of DNA sequences and its applications. *MATCH Commun. Math. Comput. Chem.* **2009**, *61*, 767–780.
75. Li, T.; Li, M.; Wu, Y.; Li, Y. Visualization methods for DNA sequences: A review and prospects. *Biomolecules* **2024**, *14*, 1447. [[CrossRef](#)] [[PubMed](#)]
76. Swetz, F.J. Leibniz, the *Yijing*, and the religious conversion of the Chinese. *Math. Mag.* **2003**, *76*, 276–291. [[CrossRef](#)]
77. Pauli, W. The influence of archetypal ideas on the scientific theories of Kepler. In *The Interpretation of Nature and the Psyche*; Jung, C.G., Pauli, W., Eds.; Pantheon Books: New York, NY, USA, 1955; pp. 147–240.
78. Ryan, J.A. Leibniz' Binary system and Shao Yong's "Yijing". *Philos. East West* **1996**, *46*, 59–90. [[CrossRef](#)]
79. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **1981**, *17*, 368–376. [[CrossRef](#)]
80. Katoh, K.; Misawa, K.; Kuma, K.I.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066. [[CrossRef](#)]
81. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **2019**, *20*, 1160–1166. [[CrossRef](#)]
82. Jung, C.G. Approaching the unconscious. In *Man and His Symbols*; Jung, C.G., von Franz, M.L., Eds.; Aldus: London, UK, 1964; pp. 18–103.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.