Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Biomedical Informatics 39 (2006) 18–33
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Homology assessment and molecular sequence alignment

Aloysius J. Phillips*

Department of Biological Sciences, Columbia University, New York, NY 10027, USA

Received 29 July 2005
Available online 9 December 2005

Abstract

Hypotheses of homology are the basis of phylogenetic analysis. All character data are considered to be equivalent regardless of the source of those characters. Putative homology statements are designated based on observations of similarity. Pairwise sequence alignment using the Needleman–Wunsch algorithm is the basis for similarity maximization between molecular sequences. Multiple sequence alignment uses this algorithm in a topologically hierarchical framework. The resulting hypotheses of homology are tested in conjunction with character congruence through parsimony. This review introduces some underlying principles of phylogenetic analysis as they pertain homology testing and DNA sequence alignment.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Homology; Cladistics; Sequence alignment; Total evidence; Simultaneous analysis; Phylogenetics

1. Introduction

Systematics is based on the notion of homology. Observed traits or characters from a set of species are homologous if they were derived from the "same" trait in the common ancestor of those species. A character is defined as any feature from an assemblage of organisms that can be evaluated as a variable with two or more mutually exclusive states. These characters may include but are not limited to morphological, genetic, or behavioral data. Analogy is the converse of homology in which characters that appear similar have evolved convergently from ancestral characters that are unrelated. Phylogenetic analysis orders species into internested sets based upon the patterns of change among the characters. These internested sets are represented as a bifurcating network called a cladogram that is rooted at its most ancestral position (Fig. 1). The principal methodological problem is how does one recognize the "sameness" of the characters if they are apt to change over evolutionary time?

There is an adage that the condition of homology is like that of pregnancy. One can be pregnant but not appear so and one can appear pregnant yet not be. Finally, one cannot be 50% pregnant. A phylogenetic character may be similar but not homologous and a character may be very dissimilar and yet be homologous. Furthermore, a character either is or is not homologous; there is no statistical element to homology. It is the historical relationship that matters. Systematics is an historical science with distinct epistemological constraints. The data is acquired through observation and not through experimental manipulation. The results of a phylogenetic analysis are only provisionally accepted pending the next new set of observations. The past can never be truly known and we can only rely on our best estimates. History has occurred only once and unique serendipitous events are pivotal during the process of evolution. If you could rewind the tape of time and replay it, then you would observe a different series of events every time you watch it. The results of a phylogenetic analysis are explicitly uncertain; accuracy is a pipe dream [1].

The concept of homology arose without regard to species evolution and natural selection. Richard Owen introduced the term in 1843 to express similarities in basic structure found between organs of animals that he considered to be more fundamentally similar than others [2]. The

* Fax: +1 212 769 5277.
E-mail address: Phillips@amnh.org.

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.11.005

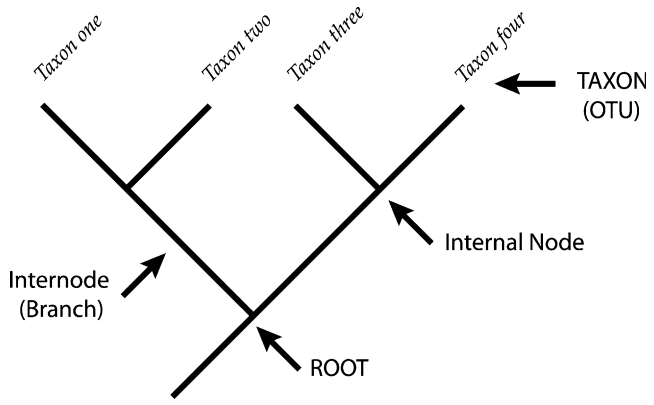

Fig. 1. Hypothetical cladogram. Descriptive terms: ROOT—the most ancestral point on the cladogram; Internal Node—the point on a cladogram where three or more branches meet; OTU—operational taxonomic unit.

term homology is derived from the Greek *homologia*, which means agreement. Owen's concept of homology was derived from Platonic and Aristotelian doctrine; there was an abstract primeval archetype upon which groups of similar animals were formed. Subsequent to Darwin, the idea of descent with modification had little impact on how homology was assessed. It did however change the mode by which homologous characters were thought to be generated, similarity due to common descent.

Other than a simple definition, what are the actual standards used to determine whether two divergent structures are or are not homologous. Remane [3] provided a set of homology criteria. Two structures are considered homologous if they fulfill one or more of the following conditions:

1. The structures are in equivalent positions within the general ground plan or organization of the organism.
2. Equivalent special quality indicated by similar cellular and tissue structure, greater complexity increases confidence.
3. Connection of differing structures by intermediate form, either developmental forms in the same organism or intermediates in different organisms.

These conventions help summarize the process of deciding whether or not a feature is comparable. Each criterion is in essence a different account of similarity. We recognize potential homologues through similarity. In practice, characters and their states are postulated as homologous on the basis of their structural, positional, ontogenetic, compositional, and/or functional correspondences.

The primary act in systematics is character analysis. Features are decomposed into their constituent parts and they are compared in terms of their positional and connectional relationships (i.e., topology) [4]. Deciding what represents a character can be problematic. Positional and structural similarity of complex structures has been taken as good evidence for homology [4]. Inevitably, there is an arbitrary component to character selection [5]. Even so, the delimitation of our empirical observations into characters with mutually exclusive character states has been effective in establishing phylogenetic relationships.

An hypothesis of character homology has been proposed as a two-step process [6]. The first step is to demarcate the boundaries of the character; what is it that is being compared? Postulates of homology through character coding are termed "primary" homologies. These postulates are then subjected to phylogenetic analysis, if the ordered series of character-state changes (transformation series) are in accord with the most parsimonious topology, then the homology assessment is accepted, this is termed "secondary" homology. If the character-state transformation series is in conflict with the topology, then the character is reassessed. Based on the judgment of the investigator, the character may or may not be recoded. Character-state delimitation and phylogenetic analysis thus have a cyclical association [7].

Brower and Schawaroch [8] prefer to reserve the term homology for characters that have been submitted to a cladistic test. They divide DePinna's primary homology into two parts. Character identification is termed "topographical identity." The second step is to assign "character-state identity." The character is defined and then discrete character states are erected.

Our concept of homology is not simply derived from our ability to ascribe similarity. Homology is a process theory necessitated by descent with modification. Phylogenetic analysis via parsimony both requires and substantiates our hypotheses of homology through mutual corroboration of the characters. In the absence of a phylogenetic hypothesis, there is no test of homology. Consequently, it is important to take into account how we execute our phylogenetic analysis and our justifications for doing so.

Homology as applied to nucleotide sequence data has had to grapple with concepts previously unknown to morphology such as gene duplication and loss, exon shuffling, and horizontal gene transfer. Nevertheless, there is no fundamental difference between the homology of molecular and non-molecular data and our analysis should reflect that. The difference lies in the mode by which we identify them as characters. For DNA sequence data, this means that we algorithmically manipulate individual strings of sequences. This review attempts to provide a brief background in phylogenetic analysis and homology of all character data as a precondition to DNA sequence alignment. Furthermore, we present the Needleman–Wunsch (NW) algorithm in the context of progressive sequence alignment for the purpose of generating phylogenetic characters.

2. Nucleotide sequence data

2.1. Orthology and paralogy

To classify different modes of homology with regard to DNA sequence data, Fitch [9] introduced the terms "orthology" and "paralogy." Orthologous sequences in two

organisms are homologs that evolved from the same sequence in their last common ancestor. Orthologs are generated as a result of speciation events. Paralogs are generated as a result of genomic duplication events; this generates two copies of the same sequence in the same lineage. This redundancy may allow one of the copies to diverge under a different set of constraints. Over the course of multiple speciation events, numerous duplications and losses can create a confusing array of orthologs and paralogs. The Hennigian auxiliary principle [7] applied to molecular data posits that characters are orthologous in the absence of any evidence indicating that they are paralogous. Nevertheless, gene phylogenies that contain paralogs do not necessarily have the same topology as the organismal phylogeny (Fig. 2). Erroneous assignment of orthology to a sequence can cause one to be misled about organismal relationships. This is one reason why it is important to look at as many different genomic regions as possible when performing phylogenetic analysis using sequence data, to ensure orthologous suites of characters.

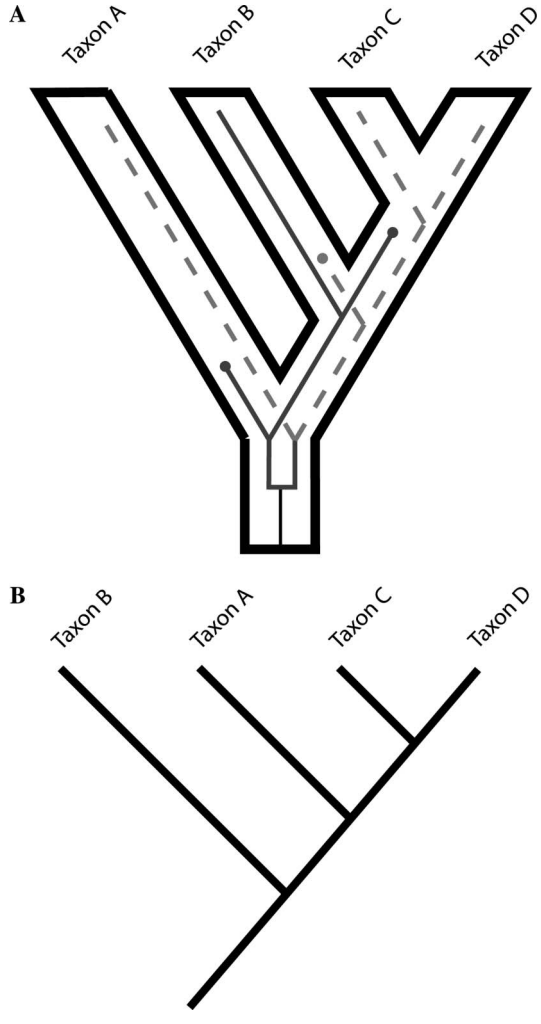

Fig. 2. An example of how gene phylogeny does not necessarily reflect the organismal phylogeny. (A) Organismal phylogeny of four taxa, embedded within the phylogeny is an example of a paralogous gene duplication in the common ancestor. Subsequent to the gene duplication, there are three gene extinction events represented as closed circles, one in the stippled lineage and two in the solid lineage. (B) The resultant gene phylogeny showing a topology that is not congruent with the organismal phylogeny.

2.2. Gene phylogenies are not necessarily species phylogenies

Phylogenetic analysis of gene sequences can be confounded by several other genetic mechanisms such as horizontal gene transfer, introgression, and ancestral polymorphism. In these instances, the gene phylogeny will not track the organismal phylogeny. Horizontal transfer is the result of an introduction into the genome across species barriers due to some vector such as a virus or transposable element. Introgression is due to transfer of genes across a reproductive barrier due to hybridization. The resultant hybrids then backcross to the original population. These alien genes can then become fixed.

The existence of polymorphisms in the ancestral population compounded with the process of lineage extinction can also yield a topology that is not representative of species relationships. When cladogenesis splits an ancestral species into two daughter species, each species can be polymorphic for the same genes. One of these species may subsequently undergo another speciation event with the result that all three species retain shared polymorphisms. When the extinction of genetic lineages occurs, the gene trees will not necessarily be congruent with the species tree. Where polymorphism occurs there is the potential for lineage sorting [10–13].

Sequencing technologies have allowed us to generate copious amounts of character data for phylogenetic analysis. More data have not necessarily transformed the logic behind which systematics is based or the way that systematic analysis is performed. An exception is parametric model-based methods such as maximum likelihood [14] and Bayesian analysis [15]. The contentions between parametric and non-parametric (parsimony) methods will not be elaborated upon here. We are generally concerned with parsimony methods as it is founded upon the concept of homology and it is the platform upon which all parametric methods are based. Parametric methods do not have a discreet homology concept. Until recently, parametric methods were only conducted with sequence data; however, all that was required to implement parametric evolutionary models upon discrete morphological data was a lack of apprehension about one's assumptions (see [16]). Parsimony is able to operate on all forms of character data with a minimum of assumptions; all characters are free to vary independently without a prerequisite of various generalized mechanisms. As will be discussed later, this is advantageous because it presents a more exacting test of homology statements.

3. Phylogenetic analysis

The test of homology is ultimately arbitrated by the character-state transformations on an hypothetical

evolutionary topology. Congruence among patterns of state change along the topology of the tree can corroborate or contradict hypotheses of homology. To understand our test of homology, it is necessary to examine the system of phylogenetic analysis.

3.1. Parsimony, homoplasy, and the auxiliary principle

The methodology of parsimony analysis, also known as cladistics, is based upon the work of Hennig [7]. He argued that only shared derived character-state changes (synapomorphies) provide information about evolutionary relationships. Characters that do not change their state and character-state changes that are uniquely derived (only present in only one species) are not informative toward phylogenetic associations. A bifurcating hierarchical network called a cladogram (Fig. 1) is constructed indicating relative relationships. The cladogram that maximizes the degree of shared derived characters (synapomorphies) among the species is the one that is chosen as the best. The principle behind this approach is the rule of parsimony; an hypothesis that requires fewer assumptions is favored over one that requires more. In this case, the assumptions are character-state transformations on the cladogram.

There is a principal supposition that there is hierarchy in the data that is representative of historical association. This hierarchy cannot be recovered without an optimality criterion otherwise all hypotheses are equal. Character conflict (homoplasy) is likely to be present to one degree or another in all datasets. The principle of parsimony as an optimality criterion minimizes this character conflict. Character-state change does not operate by any one optimality criterion; there are no universal evolutionary laws such as there are in physics (i.e., the Gas Laws). The argument that evolution is not parsimonious by extension negates all justifications of any optimality criterion in phylogenetic analysis and so is not a reasonable criticism.

Why is it necessary to minimize ad hoc arguments of homoplasy when evaluating cladograms? Scientific theories are chosen for their ability to explain our observations. The choice among these hypotheses is decided by evidence, the effect of an ad hoc explanation is to eliminate or reduce the role of an observation as evidence against a theory. Otherwise a theory need not conform to observation for it is immune to falsification [17]. When confronted with a set of theories that explain the data equally well, the simplest theory is the one with greatest explanatory power; it requires the least extrinsic information. Pliable hypotheses simply reassert the evidence; they have descriptive power but no explanatory power. Simplicity is a non-evidential constraint of phylogenetic analysis, however, this does not imply that simplicity is a characteristic of the phenomena. Hennig's (1966) system of analysis is thus underpinned by his auxiliary principle, which states that homology should be presumed in the absence of evidence to the contrary. Without the auxiliary principle all phylogenetic hypotheses are equivalent.

The entities in an analysis are usually species but they can also be higher-level groups such as genera. The general term applied to the terminal elements in a cladogram are operational taxonomic units (OTU). In reference to molecular data, OTUs can also be regions of DNA such as genes or exons especially when one is concerned with paralogous events. Paralogous gene phylogenies provide the means for studying the birth of new genes and gene functions from preexisting genes. The origin of a new gene by genomic duplication provides the raw material for molecular innovation. Paralogous gene phylogenies that are not presented in conjunction (embedded within) with organismal phylogenies are less informative because of the relative nodal information provided by both phylogenies [18].

3.2. The matrix

Once observations are established as characters and character states are assigned, the character data are represented as a rectangular matrix. The data matrix can be viewed as a set of primary homology statements. The rows usually correspond to the OTUs and the columns contain the characters; each element in a column represents a character state. For molecular data sequence alignment performs this task. Molecular data are usually encoded as an unordered multistate character. This means that there are more than two states and each state is permitted to directly change into any other state (Fig. 4A′), this is not always the case in non-molecular data.

3.3. Tree search

Parsimony algorithms do not directly generate a tree from the data. A preliminary unrooted topology (network) is chosen and the data is fit onto the network in the most parsimonious manner. Theoretically, all possible strictly bifurcating networks are then enumerated and the one(s) that require the fewest transformations to fit the data are chosen as the best (i.e., the shortest tree). In other words, the tree topology is the hypothesis that is tested by the optimization of the data. The greater the number of characters in the analysis the more stringent the test of the topology. Since there is an astronomical number of different trees to compare [32], it is computationally impossible to enumerate them all and a heuristic solution is chosen. Most of the recent advances in phylogenetic analysis involves improvements to the heuristic methods of tree search that allow a more efficient assessment of the topologies. The final step is to root the network; this identifies the most ancient node on the tree and acts to polarize the character-state transformation series. The decision of where to root the network is based on prior knowledge. The placement of the root will effect your secondary homology statements by establishing the ancestral states and determining the direction of transformations.

3.4. Congruence

A synapomorphy is a character state shared by two taxa or lineages that is derived from a singular transformation event in their common ancestor (Figs. 3A′ and B′). The two taxa constitute sister taxa. All synapomorphies are homologs. In a parsimony analysis, hypotheses of synapomorphy are tested against other hypotheses of synapomorphy. If two synapomorphy schemes suggest conflicting sets of sister taxa (taxon A, taxon B) versus (taxon A, taxon C) (Fig. 3) at least one of the schemes is wrong. At least one of the synapomorphies must be reinterpreted as an ad hoc homoplasy, i.e., a reversal or convergence (Fig. 3C′). The test of congruence is an agreement between synapomorphy schemes. Since the topology determines the degree of congruence, how we establish the topology inevitably arbitrates between what we consider to be homologous and non-homologous characters.

3.5. Transformation series

In the past, some workers encoded their morphological characters into constrained series of a priori transformations that may be linear or tree-like (Fig. 4B′), they may also have been polarized indicating the most ancestral state. These were considered hypothetical and could be recoded if they conflict with character-state transformation in the resultant analysis. However, recoding data that are

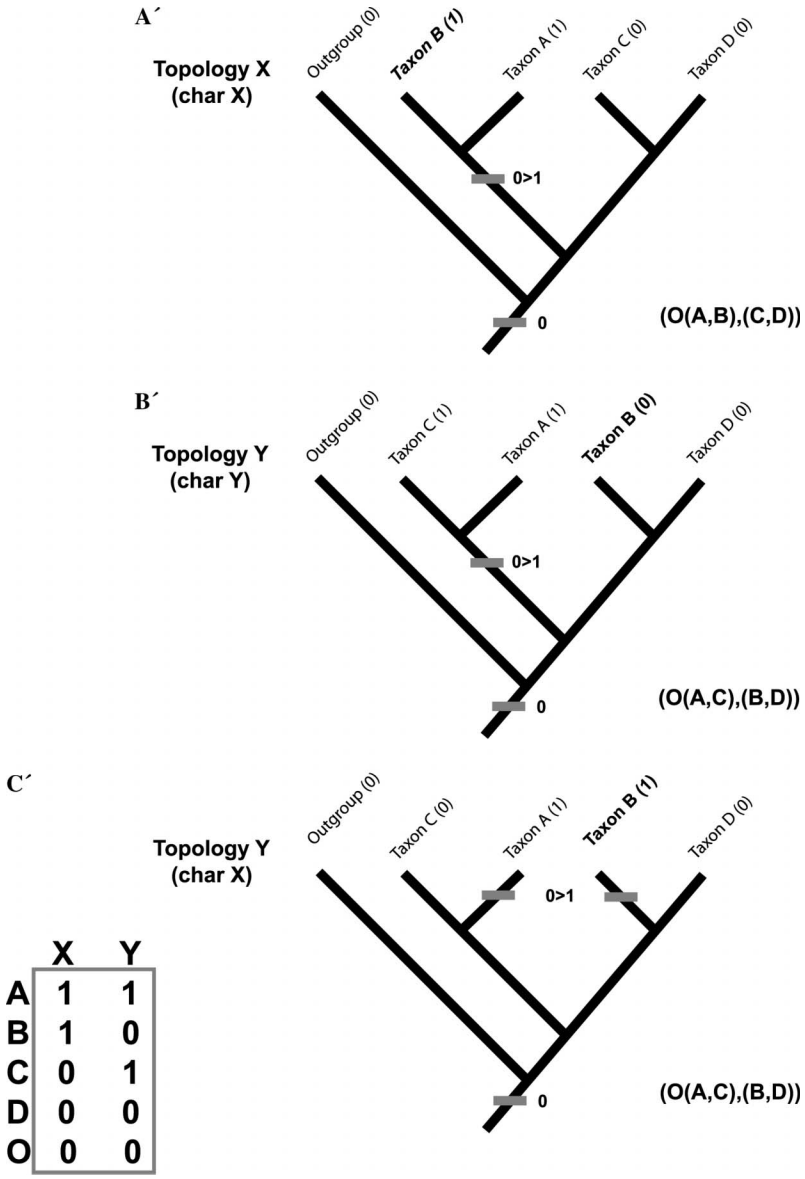

Fig. 3. An example of two characters that are incongruent with each other. (A′) When character X is optimized onto topology **X**, it requires a single transformation and is a synapomorphy for clade (A, B). (B′) When character Y is optimized onto topology **Y**, it requires a single transformation and is a synapomorphy defining clade (A, C). (C′) When character X is optimized onto topology **Y**, it requires two separate transformations. Character X is homoplasious for topology Y.

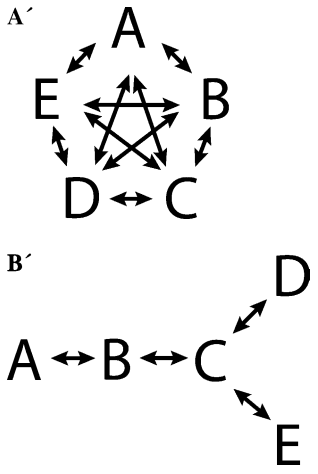

Fig. 4. Two different modes of multistate character transformation. (A′) Unordered—unordered characters can change from any state directly into any other. This type of coding requires the fewest set of assumptions. (B′) Ordered—a multistate character that has a transformation topology; transformation between non-adjacent states costs the sum of the intervening transformations. This type of coding requires special justification.

incongruent until they are congruent without justification is a dubious endeavor. Also, since the characters are attached to organisms, these a priori transformation series amount to mini-phylogenies. If these mini-phylogenies were correct, then you would expect them to be congruent with the rest of the phylogenetic signal in the absence of an a priori constraint of character change. If they are incongruent they add length (homoplasy) to the tree, which is not desirable. Why constrain the data to change in a certain way? In the absence of knowing we can only trust our data (auxiliary principle). This procedure is not commonly executed today and may be a holdover from the time when computers were not used for phylogenetic analysis and such procedures were a computational necessity. It is more parsimonious to let all the data, molecular or not, be unordered and allow the overall phylogenetic signal dictate the transformation series as a result of the analysis [19]. A character-state transformation series should be the end result of the phylogenetic analysis as it is your statement of secondary homology. To impose a priori models of character-state change obscures the test of congruence.

3.6. Homoplasy: multiple origins of the same state; is it noise or data?

The term homoplasy is used equivocally in the literature. Homoplasy is often equated with an erroneous assignment of primary homology. The characters are not the "same" and so character-state changes on the tree are ahistorical. Another explanation of homoplasy is multiple character-state changes along the topology of the tree observed as either multiple reversals or convergences; this is not necessarily explained by false character homology. It may imply incorrect character-state assignment or it may simply be the way the characters evolved. Both concepts of homoplasy are only observed as extra steps in a character-state transformation series, it is difficult to assess which type of homoplasy is the basis of those extra steps. Reversals of state may be a legitimate source of hierarchical information, whereas false homologies are not. This bias of usage may be due to the complexity of the characters that one works with. People who analyze molecular data tend not to question their primary homologies. Molecular data usually only has four states (A, T, G, and C) and so reversals and convergences may be more anticipated. People who analyze morphological data tend to expend more effort looking at each individual character and each character has a larger degree of complexity than nucleotides. Many morphological characters are rejected a priori, these suites of rejected characters are rarely considered for discussion. Observations of reversal and convergences in non-molecular characters may be more readily explained by erroneous primary homology. It is important to note that the distinction between these two kinds of homoplasy is not always made explicit.

With sequence data, a reversal can be interpreted as the creation of a non-homologous character state. In the transformation series A → T → A*, the A's are not homologous states. They are chemically indistinguishable yet have different historical identities. If you recode these "A"s as two different character states, some problems can arise because your justification for recoding is the previous phylogenetic topology and the character-state transformation series on that topology. If you recode what you observe as identical states and then reanalyze the data you abandon your justification for recoding because the resultant topology may be different and not necessarily imply a reversal. Much of the homoplasy of a dataset can be eliminated in this way. Operationally, recoding of this type of event is not attempted in a cladistic analysis.

As parsimony trees minimize homoplasy, some assert that parsimony will perform poorly if there is an abundance of homoplasy. There is no assumption that homoplasy is uncommon [17]. It is not even necessary to assume that the frequency of homoplastic events is less than synapomorphic ones. Farris [17] exemplified this point with an analogy to linear regression; although residual variance in a least-squares regression is minimized, there is no assumption that the residual variance is small. No amount of homoplasy is sufficient to discard the most parsimonious tree, you may not have a lot of confidence in it but it is the best you have with the evidence at hand. Parsimony minimizes ad hoc assumptions of homoplasy but it does not assume that homoplasy is minimal [17].

3.7. Character independence

It is a basic assumption in phylogenetic analysis that the characters are independent of one another. For example, if "number of digits on the hand" is a character, the number of digits on the left hand is not independent of the number

of digits on the right hand. If one were to encode the state of the right hand and left hand as separate characters they would be non-independent. If these two characters were congruent with the rest of the data, you would have a false degree of support. Mutually dependent characters do not reflect factual agreement and the apparent congruence is fallacious. Moreover, if the characters were homoplasious, then one would have to erect two ad hoc homoplasies to explain the fit onto the tree. This adds length to the tree, since the optimality criterion is tree length; non-independence of characters is contrary to the method. Any character-state transformation series must be logically independent from every other transformation series.

It is difficult to know beforehand whether the characters are independent or not. DePinna notes:

> The decision whether any two or more attributes comprise a single transformation series or two or more independent series is one of the most basic, albeit still confusing, issues in systematics. It is a decision that is made very early in any character analysis, and rarely questioned subsequently [6].

It is reasonable to assume that many, if not most, datasets are contaminated with non-independent characters. If the character evidence is not robust, then non-independence will be problematic. A worst-case scenario is that a homoplasious suite(s) of interdependent characters override the underlying phylogenetic signal. This form of homoplasy is non-random and more problematic than simple noise [20]. Also, measures of support can be misleading since they also assume independence. They will be inflated if the non-independent characters are congruent with the underlying signal or diminished, if they are homoplasious. As one includes a greater scope of character types, there is an implicit assumption that the characters are more likely to be independent [19]. Responsible mechanisms for non-independence include genetic linkage, pleiotropy, and ontogenetic factors such allometry and peadomorphosis.

Nucleotides that are genetically linked along a chromosome are not necessarily independent. Any selection event impinging upon a certain region of the chromosome will impact the fixation or elimination of nucleotide variants in neighboring regions. Only a crossover event inbetween any two mutations will cause them to be uncorrelated. Variation that is very far apart on the chromosome will behave independently. The closer that two characters are along the chromosome, the greater the likelihood that they will co-occur. Nucleotides within a particular gene are not independent since they comprise a functional unit; considering the functional constraints of secondary, tertiary, and quaternary structures of the resultant gene product. The nucleotides of the codon are interdependent in that they encode an amino acid. As mentioned above, there can be strings of genomic sequences that have historical topologies that are not congruent with the organismal phylogeny. The homoplasy in these suites of characters is non-random due to genetic linkage. Congruence between characters that are not genetically linked is more compelling than congruence between nucleotides that are genetically linked.

3.8. Weighting

Differential weighting of the character data is a response to homoplasy in the dataset; in the absence of homoplasy it would not be attempted. Weighting can be applied between (character weighting) or within characters (character-state weighting). Character weighting reduces the value of certain characters to the promotion of others. The decision to declare a character of weight (X) is to proceed as if the data included X independent characters all showing the same distribution of states. Character-state weighting can be represented by a stepmatrix (Fig. 5), which assigns a cost of transition from one state to another. The justification of all forms of weighting requires the burden of additional background knowledge to be instituted. Generally, the data are adjusted either before the analysis (a priori) or it is modified as a consequence of the analysis (a posteriori).

Character weighting presupposes that some characters are phylogenetically more informative than others. Even if this is true, it is difficult to know which characters are the better ones. Homoplasy (a posteriori weighting) has been used to determine a character's worth [21, 22]. A character is down-weighted in direct relation to the number of extra steps it shows on the topology. It does not necessarily follow that more homoplasious characters are less reliable as indicators of phylogeny [20,23]. Treating the frequency of events both within and between characters as a criterion to determine what is a ''bad'' character or to assign a rate expectation is incompatible with the fact that independently evolved states are historically unique [1,24]. If the homoplasy is part of the character-state transformation pattern due to reversals, then differential weighting is not justified because reversals are historical events and are informative

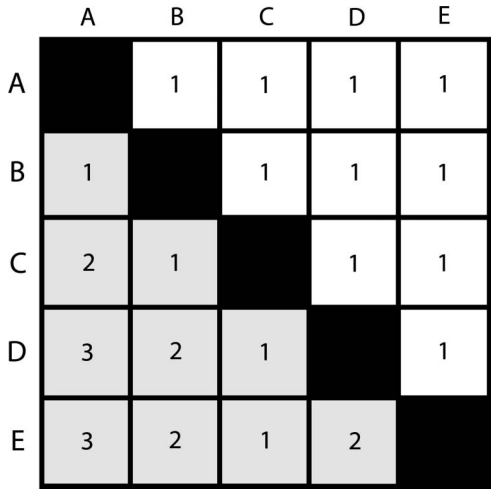

Fig. 5. A stepmatrix is a codification of the character transformation model. Above the diagonal axis is an unordered stepmatrix, below the diagonal is the stepmatrix defining the transformation series in Fig. 4B′.

toward the hierarchy. Empirical analysis of homoplasious characters in specific datasets shows that they contribute to the resolution and robustness of the phylogenies [23,25–30].

There is a spectrum of opinions about whether or to what degree one should weight. At one extreme, all characters are treated as though they have an equal potential to contribute. At the other extreme the data are compartmentalized a priori into process partitions and evolutionary models are applied, the values of the parameters of the model may be decided upon by the topology of the tree. The underlying assumption is that the extant distribution of the data may be misleading and may need to be corrected. In other words, if certain data does not fit well with the model, that data's contribution to the hypothesis is modified. Parsimony analyses tend to consider larger matrices, the underlying assumption being that more data are better and that the underlying signal(s) should be permitted to reveal themselves without undue influence from the investigator. In effect, weighting is a quantitation of your confidence in your homology assessment.

Weighting all characters equally has been said to be as arbitrary and to make as many assumptions about process as any other weighting scheme [31]. Swofford et al.'s justification for weighting is derived from a priori notions of process. Contrary to their example, empirical observation of the total percentages of different nucleotides in the matrix does not inform one about the potential behavior of any one single character. This position ignores homology and the role of topology. Frequency weighting implies that rate classes exist and that they are interdependent. If characters are independent, then the behavior of other characters should not influence them [17,24]. Equal weighting is the least arbitrary weighting scheme; it has the least assumptions of process and implies the minimum number of evolutionary transformations. The most parsimonious tree is chosen because it has the least burden of assumptions. To choose a longer tree would require the imposition of ad hoc arguments. Equal weighting also has the least burden of assumptions and so is the most parsimonious weighting scheme. This is not to say that certain weighting schemes are not valid. If one were to use codons as a character with 64 states instead of single nucleotides with only 4 states, a valid character-state weighting scheme would be to use the edit distance between each codon. This represents the minimum number of mutations required to convert one codon into another. The background knowledge would be the genetic code and the rules of translation. This example may seem trivial but that is the point, less restrained weighting schemes should not be trusted.

3.9. Simultaneous versus separate analysis

There has been a debate in systematics over how to consider datasets derived from multiple sources. The question is whether or not to analyze the data simultaneously in one grand matrix or to treat each dataset as a separate entity. Kluge [19] argues that whenever a character is defined from a group of organisms that character is assumed to contain information about the historical relationship between those organisms, regardless of the form of the character data. The organism has a unique history and the characters associated with that organism should generally reflect that. He argues that there are no discreet boundaries between different classes of character information [33]. Therefore, there is no justification for dataset separation. The total evidence analysis relies on the tenets of data independence and on the auxiliary principle. An analysis that minimizes character incongruence both within and between datasets generates the best hypothesis. Elimination or even down-weighting of data requires severe justification.

On the other side of the argument is the method termed taxonomic congruence [34,35]. This method generates a separate phylogenetic tree for each dataset and then combines the trees into a single consensus that shares the topological features of the separate trees [36]. This summary of agreement between topologies purportedly represents a conservative estimate of the phylogeny. The belief in this case is that the hypothesis is less resolved but the correct tree is embedded somewhere in the consensus. This is an attempt to minimize the effects of misleading data partitions. Parsimony is used to resolve data conflict within each dataset but there is no attempt to resolve any conflict between datasets. The attraction of taxonomic congruence is that agreement between different datasets is very unlikely due to chance, so when it is observed it instills confidence.

A more conciliatory approach is termed prior agreement [37]. In this method, it is acknowledged that simultaneous analysis of the data is desirable. However, it is asserted that datasets with a specific degree of incongruence are not combinable [38] and combinability is determined with a statistical test for heterogeneity. This incongruence is assumed to be the result of the data violating the assumptions of the method due to processes such as horizontal gene transfer; ancestral polymorphism and paralogy or the signal may also be lost due to a preponderance of homoplasy. Prior agreement does not discern what the source of the problem might be, it only rejects data combination based on heterogeneity.

Kluge's unconditional claim that there are no real boundaries between data partitions is not generally held [39]. No doubt there are a host of arbitrary partitions such as separate codon positions and the distinction between transversions and transitions, however, contiguous nucleotide sequence data such as the gene are generally accepted as a legitimate linkage partition [18]. It is at the level of the gene that most evaluations are made.

It has been claimed [40] that total evidence ignores the behavior of characters in the sense that pathological violations of assumptions are not considered. This is not exactly the case. Violations to assumptions can only be observed with reference to a phylogeny. In the absence of a phylogeny, there is no observable pattern of behavior. The assertion that patterns of character-state change are overlooked

ignores cladistic character analysis. In cladistic character analysis, characters that disagree with the bulk of the evidence are flagged as suspect; this disconfirming evidence is reexamined with regard to its primary homology. If this homoplasy is distributed non-randomly between gene regions, then a specific gene or gene region would be suspect. If the majority of the data is misleading, it is difficult to know how one would know how to correct the data and if you did try to correct the data how would you know you were right?

Consensus trees of data partitions do not directly evaluate character incongruence. Taxonomic congruence instead considers topological disagreement as a measure of character conflict. This indirect approach to character data has some drawbacks. Barrett et al. [41] found that the consensus tree from partitioned data could completely contradict the tree derived from simultaneously analyzed data. In this sense, taxonomic congruence is not conservative. Also, since the topologies are treated as equivalent in nature, if one data matrix is a different size than another, a consensus of the two trees imparts more weight to the characters of the smaller matrix. An unresolved tree requires more character-state change than a fully resolved tree. This can be misleading with respect to character evolution. In phylogenetic analysis, a less resolved tree is not the goal. Molecular sequence data are not necessarily stochastic and the distribution of both homoplasy and synapomorphy between any partitions is not necessarily heterogeneous. Also, however you choose to define a partition, there is no recourse toward internal heterogeneity of any one data partition.

The interaction of different datasets in simultaneous analysis often implies hidden character support [37,41,42]. Datasets may carry weak phylogenetic signal that is additive between partitions that overcome noise when the data are combined. Hidden support refers to the support across data partitions for relationships that are not evident in the most parsimonious tree for the partitions analyzed separately. Baker and DeSalle [26] described an index, partitioned Bremer support, to directly measure the character support from each data partition for each node of the tree from the combined data matrix. For a particular set of data partitions and for a particular group, hidden support can be defined as increased character support for the group of interest in the simultaneous analysis of all data partitions relative to the sum of support for that group in the separate analyses of each partition. The method provides a means to detect both hidden support and hidden conflict that are not apparent from separate analyses of the data. Gatesy et al. [43] have recently developed an expanded repertoire of character support indices.

Separate analysis of data partitions can only be justified as a heuristic for data exploration. The fact that hidden support exists is justification for simultaneous analysis of all data. Taxonomic congruence and prior agreement are flawed in that they both make assumptions of process prior to analysis and obscure the character evidence and so weaken the test of congruence. They substitute homoplasy with speculation of incompatible histories. No amount of homoplasy is sufficient to discard the most parsimonious solution [17]. Advocacy for a specific kind of data is not warranted. All character data have the potential to be informative toward the phylogenetic hierarchy. Sometimes when more data are added to the analysis support for the topology increases, sometimes it decreases and sometimes the topology is overturned. There is no way to predict the behavior of any character or set of characters before the phylogenetic analysis. Simultaneous analysis of all relevant data is the most justified approach because it has the fewest burden of assumptions, it is the most stringent test of homology, and it is the most informative toward the behavior of the characters and toward the phylogenetic hypothesis.

4. Multiple sequence alignment: assignment of homology to DNA sequence data

Multiple sequence alignment can have many motives. Structure prediction, motif detection, and database searching all benefit from advances in algorithmic implementation. The primary goal of these operations is to maximize some mode of similarity as rapidly as possible. However, they do not operate within an historical paradigm. Outside of the historical paradigm is not relevant whether two protein structures are the same due to common descent or through convergence. What are important are their physical properties. Most bioinformatic questions are ahistorical.

Multiple sequence alignment with the goal of phylogenetic analysis operates within an historical framework. The corresponding characters are delimited with the expectation that they contain information pertaining to the hierarchy generated through descent with modification. The alignment that is the best is not necessarily the one that maintains some physical structure or allows one to predict essential attributes. In a phylogenetic context, the best alignment is not merely the one with the shortest edit distance given specific scoring functions but also the one that implies a topology with the least internal data conflict, i.e., homoplasy. An alignment procedure embedded within cladistic parsimony can best test alternative putative homology schemes [44].

Multiple sequence alignment is the process by which one generates a phylogenetic data matrix for molecular sequence data. The columns in the matrix constitute the aligned positions in the sequences. Each aligned position is a character and the corresponding nucleotides are the states. There are three sources of ambiguity in sequence alignment, different alignment orders, parameter variation, and multiple equally costly alignments [45].

4.1. Progressive multiple sequence alignment

Hogweg and Hesper [46] and Feng and Doolittle [47] introduced the strategy of progressive alignment. Multiple

sequences are aligned sequentially in a pairwise manner. An alignment topology or guide tree is generated that directs the order by which the sequences are aligned. Typically, only one guide tree is generated. There is an order dependency by which sequences are accreted to the multiple sequence alignment [48]. If you change the topology of the guide tree, you can alter the resultant multiple alignment. Each node in the guide tree represents an alignment. As one proceeds down the tree, these nodal subalignments are combined using various criteria. Some methods use consensus sequences and so the deeper nodes in the guide tree are aligning ambiguous sequences. Other programs use profile alignments [49]. This can change the cost regime of the alignments at the internal nodes of the alignment topology. Any gaps inserted higher up in the tree are maintained through any subsequent subalignment. Programs based on this method have become very popular, chief among these is the ClustalW program [50,51].

Congruence tests your homology scheme and so it tests your alignment. If a single parameter set yields several alignments, the best alignment is the one that has the minimal amount of incongruence on the tree. That is to say, the alignment that generates the shortest tree is the best alignment given that specific parameter set [44]. If you were to take the same data matrix and weight the characters according to two different schemes, the lengths of the trees are not comparable. Each tree minimizes incongruence within its own weighting scheme. In essence your homology schemes have changed, the dataset is treated differently. It is difficult to state which of the trees is a better result. With DNA sequence alignment, the alignment parameters are tantamount to a weighting scheme. Your alignment parameters establish the homology. Therefore, the greatest test of homology for a given alignment is one that uses the alignment parameters as its weighting scheme. Otherwise, we are generating our hypotheses under one set of assumptions and testing those hypotheses under a different set of assumptions. This weakens our test of congruence. Phylogenetic analyses that have inconsistent alignment/weighting schemes and alignment programs with internally inconsistent cost schedules obscure homology assessment.

The ClustalW packages do what they do well, however, the alignment parameters change dynamically during alignment both within and between sequences. It is difficult to recover what the cost regimes were after an alignment is finished. If one wants to find the best-supported phylogenetic alignment through character congruence, the transformation costs are not easily recoverable. Even though the ClustalW algorithm is used widely in phylogenetic analyses, it is not based on a protocol that maximizes similarity in a historical framework. Alignments that maximize historical information do not necessarily maintain aesthetically pleasing structural blocks. Benchmarking alignments to conserved motifs are not justified in this case. If one states that a motif is ''conserved,'' that statement should be in a context of a phylogeny. The characters and hence the alignment have passed the test of congruence. Constraining an alignment to maintain specific structures disregards the potential for these structures to evolve. A priori the character states are ancestral; there is no test of congruence. Any historical evidence to the contrary has been removed before the phylogenetic analysis. Many ''conserved'' structures are prevalent because of some kind of constraint. We have no knowledge of whether these constraints have been persistent over evolutionary time; we do not even know what the constraints are in most cases. Constraint can maintain a structure and hence it is ancestral or it can cause convergence. Only within a phylogeny can one discern between the two scenarios.

Almost all multiple sequence alignment procedures are a series of progressive pairwise alignments. It is possible to perform simultaneous alignments of many sequences in a multidimensional matrix [52] although this is computationally intractable. Theoretically, simultaneous alignment may be a more efficient description of overall similarity, however, progressive alignment utilizes a strictly bifurcating scheme that may be considered analogous to cladogenesis. A simultaneous multiple alignment would imply a star phylogeny with no hierarchy. Since we test our primary homology through character congruence [19], our assumption sets should remain static during phylogenetic analysis. I see this as logical justification for progressive sequence alignment to establish primary homology in a hierarchical framework.

What is the correct progressive alignment topology? Given that the true phylogeny is unknowable, a true alignment topology is also not feasible. Consider that different multiple alignment topologies can yield the same alignment and so even though our alignment topology connotes the phylogeny it is not representative of it. It is possible that an alignment topology will be congruent with its resultant phylogeny but this cannot be seen as a goodness-of-fit criterion. A suboptimal tree that is congruent with its alignment topology cannot be chosen over an optimal tree that is incongruent with its alignment topology. As such, given two equally parsimonious trees, one cannot favor the tree that is most congruent with its alignment topology. That is a criterion outside the realm of homology and character congruence. The best alignment topologies are the ones that generate the alignments that yield the shortest trees.

4.2. Pairwise alignment: the Needleman–Wunsch algorithm

The primary procedure to perform pairwise sequence alignment is the NW algorithm [53]. Since this is the distinct process whereby positional primary homology is assigned it will be discussed in detail.

The NW algorithm calculates a minimum edit distance between two strings of characters. This is the minimum number of transformation operations required to convert one sequence into another. The two basic operations are gap placement and mismatches. Each operation is associated with a penalty; the sum total of these operations is the

edit distance. The penalty structure or parameter set can range from simple to very complicated. The simplest scheme is a cost of 1 for any mismatch and a cost of 1 for any gap placement. Mismatches can be separated into categories and independent costs assigned to each type of transformation. For instance, transitions can be weighted differently from transversions. In the case of amino acids, log odds matrices such as PAM [54] or Blossum [55] are often used.

The unitary or simple gap cost treats each gap placement as independent of every other gap regardless of its location. A more complicated gapping protocol is to apply affine gap costs [56–59]. Affine gap cost is determined by whether the gap characters are contiguous or not. These gapping methods treat each contiguous string of gaps as a single event. The penalties are composed of two parts, the initiation cost and the length-dependent extension cost. The extension cost is often not a linear function. Concave gapping decreases the cost of each incremental inclusion of a gap character, often as a log function [59–61]. As with any other weighting scheme, there is no one correct set of gap parameters.

The use of gaps as characters in phylogenetic analysis has been shown to be of great value [62–64]. The application of affine gap costs in phylogenetic analysis is problematic. The gap may be perceived as a single event but it is still encoded as separate gap characters in the data matrix. This will cause non-independence between gap characters [65]. It is possible to manually recode multiple residue gaps as single characters with the different character states determined by length differences and corresponding nucleotide variation [66–68]. For consistency the transformation cost between the manually recoded states should be their edit distance based on the alignment parameters.

The NW algorithm belongs to a class of algorithms called dynamic programming. The idea is to break a large problem down into incremental steps or subproblems so that, at any given stage, optimal solutions to the subproblems are known. This condition can be extended incrementally without having to alter previously computed optimal solutions to subproblems. The best solution to the global problem is found by using an optimal combination of the smaller subproblem solutions through a retroactive formula that connects the solutions. The NW algorithm proceeds as follows:

Step 1 Initialize the matrix. A pairwise matrix is constructed of $(M+1)\times(N+1)$ cells, M is the length of one sequence and N is length of the other. The first cell in the matrix (0,0) is the null cell. Each cell is filled with a mismatch penalty, in this case either 1 for a mismatch and 0 for a match (Fig. 6).

Step 2 The wavefront update of the matrix. Starting at the null cell (0,0) and proceeding toward the terminal cell [8,10], each cell in the matrix is evaluated within the context of a recursive set of overlapping subproblems. The results of which are stored in memory and are used to determine the solution of the next subproblem. Each cell calculates the least costly path or paths from each of three previous adjacent cells in the matrix (Fig. 7). The paths represent a separate edit operation. The optimal pathway or pathways are saved in memory, the sum total of the cost of the edit operations up to that point are stored in the cell (Fig. 8). This local edit distance is then used to evaluate the edit path to its adjacent cells, to the right, on the diagonal, and below. Once a cell has contributed to a subproblem, its edit distance is no longer of any value since that path through the matrix has either terminated or has been incorporated into the next edit distance operation.

Step 3 Traceback. The traceback begins at the terminal element of the matrix [8,10]. Any previously

(0,0)		T	T	T	C	C	A	A	G	G	C
	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	1	1	1	1	1	1	1
T	0	0	0	0	1	1	1	1	1	1	1
T	0	0	0	0	1	1	1	1	1	1	1
C	0	1	1	1	0	0	1	0	1	1	0
A	0	1	1	1	1	1	0	0	1	1	1
G	0	1	1	1	1	1	1	1	0	0	1
C	0	1	1	1	0	0	1	1	1	1	0
C	0	1	1	1	0	0	1	1	1	1	0

(10,8)

Fig. 6. An initialized matrix of a pairwise nucleotide sequence comparison with an assigned mismatch cost of 1.

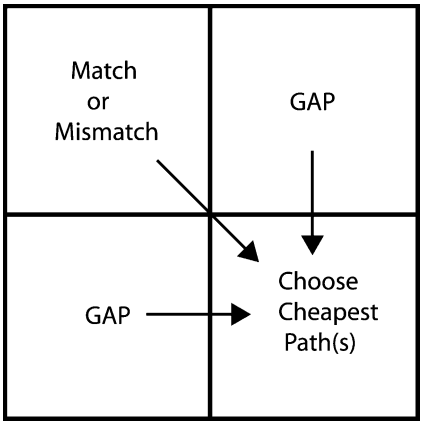

Fig. 7. The kernel of the Needleman–Wunsch dynamic programming subproblem. Every cell in the matrix must not only arbitrate the cheapest path to itself but must also contribute to the arbitration of its three potential neighbors toward the terminal cell.

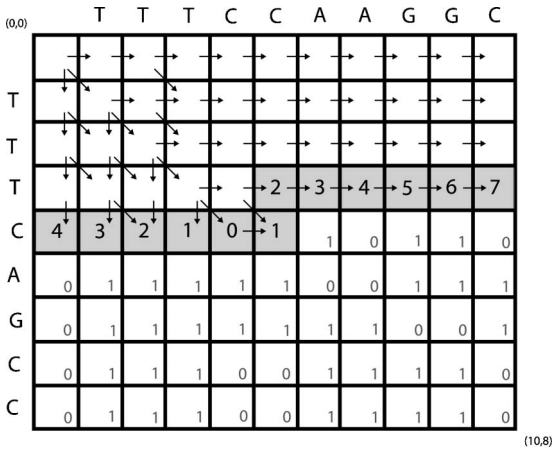

Fig. 8. The wavefront update of the matrix. The arrows indicate the optimal paths to the next cell. The highlighted cells represent the wavefront, they contain the local edit distance used in the calculation of the next round of optimal paths. These edit distances are saved in memory until a cell can no longer contribute to the optimal path calculation. The cells below the wavefront contain the mismatch cost that will be used in the calculation of the optimal path. The arbitration is as follows, a diagonal path to a cell represents a match/mismatch, if the cell contains a mismatch value of 1, then a cost of 1 is added to the edit distance of the previous cell and that new value is applied to the current cell. If a cell contains a mismatch value of 0, then a match is implied and there is no cost added to the edit distance, the current cell will be assigned the same value as the previous cell. Gaps are implied through edit paths that are not on the diagonal, in this case, they are always assigned a cost of 1. Since gaps are neither matches nor mismatches, the mismatch cost is not applied. The current cell obtains the edit distance of the previous cell plus the gap cost. The kernel is applied recursively until the terminal cell is reached.

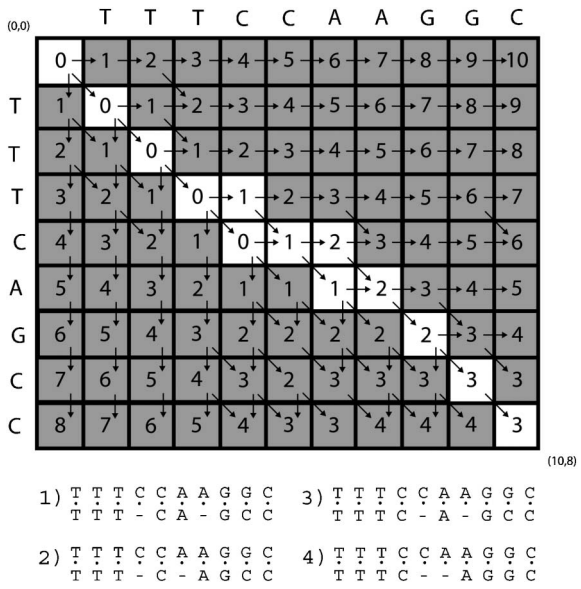

Fig. 9. The traceback. Beginning at the terminal cell, all paths into that cell are followed. All possible contiguous paths through the matrix are followed back to the null cell (0,0). These paths represent the edit graphs. In this example, there are four equally costly edit graphs. These represent four equivalent primary homology hypotheses.

retained optimal path to the terminal element is followed to the cell that is the source of that path. This process is repeated recursively and a contiguous path of optimal subproblem solutions is traced through the matrix toward the null cell (0,0). If two paths were retained in memory, this spawns a branchpoint in the traceback creating multiple paths through the matrix. This path or paths represents the edit graph between the two sequences. This alignment represents the primary homology assignment of the sequence data. In our example, there are four different alignments (Fig. 9) each of which is an equally likely homology hypothesis.

The optimal path through the pairwise alignment matrix is sensitive to variations in the cost functions. Thus, our hypothesis of primary homology is determined by the cost functions in the alignment. In essence the cost functions are an a priori weighting scheme. Any alignment generated by a different set of parameters yields a different set of homology statements; it is a different phylogenetic dataset. The choice of the alignment parameters is largely arbitrary although it is logically bounded by the triangle inequality [69]. A gap cost of 0 yields an alignment with no mismatches with an edit distance of 0; this is not a valid parameter choice.

As shown above, there can be more than one optimal pairwise alignment for any one set of alignment parameters. Most of the common alignment programs will only report one of many possible alignments. Typically, when confronted with a branchpoint a program will have an embedded decision process. It may consistently choose the diagonal path (a mismatch) over the horizontal or vertical paths (gaps) or when given a choice between gaps, it will take only the horizontal path or only the vertical path. This in essence biases all gaps either to the 5′ end of the sequences or the 3′ end. If you repeat the alignment you will get the same result, this may give a false sense of consistency to the uninitiated user. It also surreptitiously obscures homology statements. The alignment is the data; it is not advisable to systematically bias the data by occluding alternate alignments. Equivocal hypotheses of homology should be arbitrated through congruence. Scrutiny of the above procedure will show that each column in the aligned data matrix has an algorithmic dependency based on a distance metric to every other column in the matrix. This indicates another degree of non-independence of linked nucleotide sequence data [65,70].

4.3. Malign

Most multiple alignment packages use a distance-based method to generate a single guide tree topology. The justification is that the two most similar sequences are from sister taxa. This is wrong [17]. Malign is an alignment program [71] that considers all possible multiple alignments topologies (or a heuristic thereof) and chooses the alignment topology that provides the most parsimonious tree.

The alignment parameters are maintained as character-state weighting schemes in the phylogenetic analysis. If several guide trees are found to be equally optimal, Malign can output all resultant matrices. Malign is also capable of reporting all equally optimal alignments of each pairwise alignment. In this, Malign is unique among multiple sequence alignment programs.

Malign will also institute a protocol called elision [72]. Elision is a method that concatenates the results of alignments generated with different cost regimes. The grand alignment is then analyzed. The desired effect is that congruent alignment variable positions will contribute additively to the phylogenetic signal, whereas homoplasious alignment variable positions will be non-additive noise. This method can also be used to explore multiple alignment paths. Malign produces multiple sequence alignments solely in the context of a phylogeny.

4.4. Suboptimal alignments

All paths through the pairwise matrix represent an alignment. Most of these are suboptimal edit paths. All possible paths can be enumerated and ranked according to edit distance. If the best alignment is that which provides the most parsimonious solution, then why not also explore a set of suboptimal alignments? There is no a priori reason to assume that the shortest edit distance is the best hypothesis of homology. The edit distance is a similarity metric and not a test of homology. One issue with using suboptimal alignments is that there is no good stopping rule. How many extra steps in the edit distance do we explore? Including more sequences compounds an already intractable computational problem. But is that a justification not to explore more ground for enhanced homology assessments? One predicament is that as the alignment process becomes less constrained it can generate datasets with fewer phylogenetically informative sites. These data may generate phylogenies that are less costly and have very little internal data conflict. If suboptimal alignments are used as data, the phylogenetic tree length can no longer be used to arbitrate between alignments. What is the criterion for deciding when a suboptimal alignment has become a poor estimation of homology? It is a slippery slope to abandon the simplest alignment solution.

4.5. Dynamic sequence homology

In traditional systematic analysis, the data matrix is static during the tree search. If tree search and sequence alignment are performed simultaneously, this need not be the case. Optimization alignment [73] as implemented in the program POY [74] optimizes the total character transformation cost of a set of sequences over a topology but without establishing conventional multiple alignment matrices. Only pairwise comparisons are considered at the nodes of the network. Gaps in an alignment are not observations; they are constructs [73]. In optimization alignment as in real sequences, there are no gaps only insertion deletion events. Primary homology schemes are in flux because the phylogenetic topology and the alignment topology are the same. As the topology changes so too can the primary homologies. This method is able to simultaneously consider static matrices such as those used in morphology. In a simultaneous analysis framework, the morphology (including fossil data) matrix will influence the dynamic alignment and the dynamic alignment will influence the character-state transformation series of the static matrix. This may seem absurd from the perspective of ahistorical similarity maximization but in the context of global congruence between all the phylogenetic characters it is sound. If the data is independent, then there is no justification to exclude one type of data from the analysis, this includes multiple gene regions and non-molecular data. The optimality criterion is the shortest length tree. This is the tree with the most congruence among all the data points. This is the tree that has undergone the most stringent test of homology. The best phylogenetic alignment is the one that agrees with all of the relevant data, this includes other alignments and non-molecular data.

4.6. Genomic strings as characters

The ideal unit of phylogenetic analysis is one that evolves independently of other units. They represent separate pieces of evidence. This is not necessarily a sound assumption for molecular sequence data. The effect of genetic linkage, being a component of a larger functional unit i.e., the gene, and the algorithmic artifact of pairwise alignment render the assumption of independence among nucleotide sequence data suspect. There is no a priori justification for single nucleotides being the fundamental phylogenetic unit character other than that they are irreducible. The notion of strings of sequence data being the comparable homologous units and not the individual nucleotides or amino acids is not novel [13,75,76]. Albert et al. [75] have argued for strings as characters from the perspective that highly complex characters are less apt to be effected by reversals of state. An advantage to strings as characters approach is that it decreases the probability of independent gains of the same character state. Genes as characters have a greater character-state space and have a much-reduced likelihood of reversals of state, whereas single nucleotides have a much greater likelihood of reversal of state. Genes that do not track the phylogeny of the organism due to violations of assumptions would be observed as single homoplasious characters. Semple and Steel [77] suggest that it may be possible to reconstruct large trees with only a small number of these complex characters perhaps on the order of five as opposed to thousands of characters as required for nucleotides.

Strings as complex characters may ameliorate many of the above-mentioned problems with non-independence of linked characters. In the fixed states optimization method of analysis [78] implemented in POY [70], each sequence

is treated as a character state in a step matrix (Fig. 5). The transformation cost between states is determined by its edit distance. In a scheme such as this, the alignment parameters determine the transformation cost. So, again the pairwise alignment parameters affect your homology schemes, but in this case, it is transformation cost and not the assignment of aligned positions. The most parsimonious tree is essentially the least costly set of character-state transformation series. Complex gapping protocols are unambiguously accommodated in this paradigm. Instead of a string of non-independent gap characters in a static matrix the gap cost is incorporated into the overall transformation cost. By extension, this mode of analysis can be applied to entire genomes. Sequence transformations would not only include mismatches and gaps but also translocations, duplications, and inversions. The entire genome could be a single character.

5. Conclusions

Homologous characters retain evidence of phylogenetic history. We erect theories of homology based on observations of similarity. Which form that similarity takes, structural, ultrastructural, positional, developmental, even functional, is not relevant. One cannot know beforehand whether sets of similar characters are in fact homologous. A phylogenetic character may be similar but not homologous and a character may be very dissimilar and yet be homologous. The phylogenetic tree will illustrate whether the characters arose independent of history.

Character delimitation is fundamental to phylogenetic analysis. Sequence alignment is the process by which we assign putative homology to molecular data. The NW algorithm simply maximizes similarity within a specific cost regime, but this is the recognition criterion of homology. Too often the process of alignment is considered incidental even though it is the operation that generates the dataset. Since there is no difference between homology of molecular and non-molecular data, all relevant characters should be tested against one another simultaneously, this is the most severe test of congruence. In a phylogenetic context, the best alignment is the one that generates the most parsimonious tree when analyzed in conjunction with all relevant data.

Phylogenetic history is an unknowable entity. If we accept character data as evidence of the past, then the quality of our character data is also by extension unknowable because they are the products of these past events. We constrain our hypotheses of homology through parsimony. Thus we submit to the weight of the character evidence. We should be agnostic toward ideas about mechanisms prior to phylogenetic analysis. The confounding degree of organismal complexity and the role of contingency in evolution cannot be rescued by the overriding simplifications of arbitrary process models. If a phylogenetic hypothesis is inconclusive, it is far more defensible to get more data than it is to squelch the data that you do have. Upon inspection, simplistic notions of process generally fall apart. Classes of data are artificial constructs. Linkage partitions are non-independent suites of characters. Much of the observation of ''process'' is the result of violations of the underlying assumption of independence.

Recovery of conserved motifs is an ahistorical criterion for alignment quality. Constraining alignments to fit these motifs disregards potentially informative events for phylogeny. A priori ideas about structure create forbidden alignment space. This precludes numerous possible historical scenarios; many states in evolution are not necessarily optimal. Constraining alignment space in this way suffers from the same setbacks as a priori process models.

References

[1] Siddall ME, Kluge AG. Probabilism and phylogenetic inference. Cladistics 1997;13:313–36.
[2] Owen R. Lectures on the comparative anatomy and physiology of the invertebrate animals, delivered at the Royal College of Surgeons, in 1843. London: Longman, Brown, Green and Longmans; 1843.
[3] Remane A. Die grundlagen des natürlichen systems, der verleichenden anatomie und der phylogenetic. Leipzig: Akademische Verlagsgesellschaft; 1952.
[4] Rieppel O. Homology, topology, and typology: the history of modern debates. In: Rieppel Hall BK, editor. Homology: the hierarchical basis of comparative biology. San Diego: Academic Press; 1994.
[5] Pogue M, Mickevich M. Character definitions and character-state delimitations: the bete noire of phylogenetic inference. Cladistics 1990;6:319–61.
[6] de Pinna MCC. Concepts and tests of homology in the cladistic paradigm. Cladistics 1991;7:367–94.
[7] Hennig W. Phylogenetic systematics. Urbana: University of Illinois Press; 1966.
[8] Brower AVZ, Schawaroch V. Three steps of homology assessment. Cladistics 1996;12:265–72.
[9] Fitch WS. Distinguishing homologous from analogous proteins. Syst Zool 1970;19:99–113.
[10] Goodman MJ, Czelusniak GW, Moore GW, Romero-Herrera AE, Matsuda G. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 1979;28:132–63.
[11] Avise JC, Shapiro SW, Daniel SW, Aquadro CF, Lansman RA. Mitochondrial DNA differentiation during the speciation process of *Peromyscus*. Mol Biol Evol 1983;1:38–56.
[12] Pamilo P, Nei M. Relationships between gene trees and species trees. Mol Biol Evol 1988;5:568–83.
[13] Doyle J. Gene trees and species trees: molecular systematics as one-character taxonomy. Syst Bot 1992;17:144–63.
[14] Huelsenbeck JP, Crandall KA. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst 1997;28(1):437–66.
[15] Lewis PO, Swofford DL. Phylogenetic systematics turns over a new leaf. Trends Ecol Evol 2001;16(1):30–7.
[16] Lewis PO. Maximum likelihood phylogenetic inference: modeling discrete morphological characters. Syst Biol 2001;50:913–25.
[17] Farris JS. The logical basis for phylogenetic analysis. In: Farris Platnick NJ, Funk VA, editors. Advances in cladistics. New York: Columbia University Press; 1983.
[18] Page RDM, Charleston MA. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 1997;7:240–321.
[19] Kluge A. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst Zool 1989;38:24–38.

[20] Wenzel JW, Siddall ME. Noise. Cladistics 1999;15:51–64.
[21] Farris JS. A successive approximations approach to character weighting. Syst Zool 1969;18:413–44.
[22] Goloboff PA. Estimating character weights during tree search. Cladistics 1993;9:83–91.
[23] Källersjö M, Albert VA, Farris JS. Homoplasy increases phylogenetic structure. Cladistics 1999;15:91–3.
[24] Kluge A. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic analysis. Zool Scr 1998;26:349–60.
[25] Philippe H, Lecointre G, Van Le HL, Le Guyader H. A critical study of homoplasy in molecular data with the use of a morphologically based cladogram, and its consequences for character weighting. Mol Biol Evol 1996;13(9):1174–86.
[26] Baker RH, DeSalle R. Multiple sources of character information and the phylogeny of *Hawaiian drosophilids*. Syst Biol 1997;46:654–73.
[27] Olmstead RG, Reeves PA, Yen AC. Patterns of sequence evolution and implications for parsimony analysis of chloroplast DNA. In: Soltis PS, Soltis PS, Doyle JJ, editors. Molecular systematics of plants II: DNA sequencing. Boston: Kluver Press; 1998. p. 164–87.
[28] Bjorklund M. Are third positions really that bad? A test using vertebrate cytochrome *b*. Cladistics 1999;15:191–7.
[29] Sennblad B, Bremer B. Is there justification for differential a priori weighting in coding sequences? A case study from *rbc*L and Apocynaceae s.l. Syst Biol 2000;49(1):101–13.
[30] Baker RH, Wilkinson GS, DeSalle R. Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). Syst Biol 2001;50:87–105.
[31] Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Sunderland (MA): Sinauer Associates Inc.; 1996.
[32] Felsenstein J. The number of evolutionary trees. Syst Zool 1978;27:27–33.
[33] Kluge A, Wolf AJ. Cladistics: whats in a word. Cladistics 1993;9:183–99.
[34] Mickevich M. Taxonomic congruence. Syst Zool 1978;27:143–58.
[35] Miyamoto M, Fitch W. Testing species phylogenies and phylogenetic methods with congruence. Syst Biol 1995;44:64–7.
[36] Nelson G. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's *Familles des Plantes* (1763–1764). Syst Zool 1979;28:1–21.
[37] Chippindale PT, Wiens JJ. Weighting, partitioning, and combining characters in phylogenetic analysis. Syst Biol 1994;43:278–87.
[38] Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ. Partitioning and combining data in phylogenetic analysis. Syst Biol 1993;42(3):384–97.
[39] Nixon K, Carpenter JM. On simultaneous analysis. Cladistics 1996;12:221–41.
[40] Huelsenbeck JP, Bull JJ, Cunningham CW. Combining data in phylogenetic analysis. Trends Ecol Evol 1996;11:152–8.
[41] Barrett M, Donoghue MJ, Sober E. Against consensus. Syst Zool 1991;40:486–93.
[42] Olmstead RG, Sweere JA. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst Biol 1994;43(4):467–81.
[43] Gatesy J, O'Grady P, Baker R. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 1999;15:271–313.
[44] Wheeler WC. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst Biol 1995;44:321–32.
[45] Wheeler WC. Sources of ambiguity in nucleic acid sequence alignment. In: Schierwater B, Streit B, Wagner GP, DeSalle R, editors. Molecular ecology and evolution: approaches and applications. Basel: Birkhäuser Verlag; 1994. p. 323–52.
[46] Hogweg P, Hesper P. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 1984;20:175–86.
[47] Feng D-F, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 1987;25:351–60.
[48] Lake JA. The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol 1991;8:378–85.
[49] Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–8.
[50] Thompson JD, Higgins DG, Gibson TJ. Clustal-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[51] Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 1996;266:383–402.
[52] Sankoff D, Cedergren RJ. Simultaneous comparison of three or more sequences related by a tree. In: Sankoff D, Kruskal JB, editors. Time warps string edits and macromolecules: the theory and practice of sequence comparison. London: Addison-Wesley; 1983. p. 253–63.
[53] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53.
[54] Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequences and structure. Washington (DC): National Biomedical Research Foundation; 1978. p. 345–52.
[55] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1982;89:10915–9.
[56] Waterman MS, Smith TF, Beyer WA. Some biological sequence metrics. Adv Math 1976;20:376–87.
[57] Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–8.
[58] Waterman MS. General methods of sequence comparison. Bull Math Biol 1984;46:473–500.
[59] Miller W, Meyers EW. Sequence comparisons with concave weighting functions. Bull Math Biol 1988;50:97–120.
[60] Knight JR, Meyers EW. Approximate regular expression pattern-matching with concave gap penalties. Algorithmica 1995;14:85–121.
[61] Allison L. Normalization of affine gap costs used in optimal sequence alignment. J Theor Biol 1993;161:263–9.
[62] Eernisse D, Kluge AG. Taxonomic congruence versus total evidence, amniote phylogeny inferred from fossils, molecules, and morphology. Mol Biol Evol 1993;10:1170–95.
[63] Vogler AP, DeSalle R. Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. Mol Biol Evol 1994;11:393–405.
[64] Giribet G, Wheeler WC. On gaps. Mol Phylogenet Evol 1999;13(1):132–43.
[65] Phillips A, Janies D, Wheeler WC. Multiple sequence alignment in phylogenetic analysis. Mol Phylogenet Evol 2000;16(3):317–30.
[66] DeSalle R, Brower AVZ. Process partitions, congruence, and the independence of characters: inferring relationships among closely related Hawaiian *Drosophila* from multiple gene regions. Syst Biol 1997;46(4):751–64.
[67] Gan WB, Wong VY, Phillips A, Ma C, Gershon TR, Macagno ER. Cellular expression of a leech netrin suggests roles in the formation of longitudinal nerve tracts and in regional innervation of peripheral targets. J Neurobiol 1999;40(1):103–15.
[68] Simmons MP, Ochoterena H. Gaps as characters in sequence-based phylogenetic analysis. Syst Biol 2000;49(2):369–81.
[69] Wheeler WC. The triangle inequality and character analysis. Mol Biol Evol 1993;10(3):707–12.
[70] Gatesy J, Amato G, Norell M, DeSalle R, Hayashi C. Combined support for wholesale taxic atavism in gavialine crocodylians. Syst Biol 2003;52(3):403–22.
[71] Wheeler WC, Gladstein DS. MALIGN: a multiple sequence alignment program. J Hered 1994;85:417–8.

[72] Wheeler WC, Gatsey J, DeSalle R. Elision: a method for accommodating multiple molecular sequence alignments with alignment ambiguous sites. Mol Phylogenet Evol 1995;4:1–9.
[73] Wheeler WC. Optimization alignment: the end of multiple sequence alignment in phylogenetics. Cladistics 1996;12:1–9.
[74] Wheeler WC, Gladstein DS, De Laet J. POY. Phylogeny reconstruction via optimization of DNA and other data. 3.0.11 ed.; 2003. Available from: http://research.amnh.org/scicomp/projects/poy.php.
[75] Albert V, Backlund A, Bremer K, Chase M, Manhart J, Mishler B, et al. Functional constraints and *rbc*L evidence for land plant phylogeny. Ann Mo Bot Gard 1994;82:534–67.
[76] Patterson C. Homology in classical and molecular homology. Mol Biol Evol 1988;5:603–25.
[77] Semple C, Steel M. Tree reconstruction from multistate characters. Adv Appl Math 2002;28:169–84.
[78] Wheeler WC. Fixed character states and the optimization of molecular sequence data. Cladistics 1999;15:379–85.