Is Sequence Alignment an Art or a Science?

Author(s): David A. Morrison

Source: Systematic Botany, 40(1):14-26.

Published By: The American Society of Plant Taxonomists

URL: http://www.bioone.org/doi/full/10.1600/036364415X686305

Systematic Botany (2015), 40(1): pp. 14–26
© Copyright 2015 by the American Society of Plant Taxonomists
DOI 10.1600/036364415X686305
Date of publication February 12, 2015

Is Sequence Alignment an Art or a Science?

David A. Morrison

Systematic Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden
Author for correspondence (David.Morrison@ebc.uu.se)

Communicating Editor: Mark P. Simmons

Abstract—Aligning multiple nucleotide sequences is a prerequisite for many if not most comparative sequence analyses in evolutionary biology. These alignments are often recognized as representing the homology relations of the aligned nucleotides, but this is a necessary requirement only for phylogenetic analyses. Unfortunately, existing computer programs for sequence alignment are not based explicitly on detecting the homology of nucleotides, and so there is a notable gap in the existing bioinformatics repertoire. If homology is the goal, then current alignment procedures may be more art than science. To resolve this issue, I present a simple conceptual scheme relating the traditional criteria for homology to the features of nucleotide sequences. These relations can then be used as optimization criteria for nucleotide sequence alignments. I point out the way in which current computer programs for multiple sequence alignment relate to these criteria, noting that each of them usually implements only one criterion. This explains the apparent dissatisfaction with computerized sequence alignment in phylogenetics, as any program that truly tried to produce alignments based on homology would need to simultaneously optimize all of the criteria.

Keywords—Multiple alignment, alignment algorithm, sequence homology.

Multiple sequence alignment software have not yet met their primary aim for evolutionary biologists: maximizing homology of characters. This is in spite of 30 yr of work in the field by scores of people (starting with Hogweg and Hesper 1984). All of this effort has led to a proliferation of alignment methods that have diverse optimization functions, along with assorted heuristics to search for the optimum alignment. These methods produce detectably different multiple sequence alignments in almost all realistic cases, which leaves the phylogenetics practitioner wondering what to do.

If the goal is to develop an automated procedure for homology assessment, then we currently do not have one, and no one has demonstrated where we might get one in practice. It is worth looking at why, and also how we might make some progress in the near future. My purpose here is therefore to try to conceptualize why there are currently so many different approaches to sequence alignment (e.g. see the lists of programs in Do and Katoh 2009; Anisimova et al. 2010), and see how they relate to each other in the context of homology assessment.

I start by putting aside the automation issue for the moment, and looking first at the actual biological goal (nucleotide homology). I try to identify the traditional paradigm for detecting homology, and then explicitly relate this to nucleotide sequences. Only then do I consider whether / how this paradigm might be automated.

Homology as a Goal for Alignment

Homology is a topic of long-standing interest to biologists (Hall 1994; Bock and Cardew 1999; Wagner 2001; Kleisner 2007). This follows from the idea that both homologies and phylogenies need to be “discovered” within the phenotypic and genotypic data that we have accumulated about biological organisms. How do we go about this discovery?

If we accept the idea that there is no fundamental difference between homology in classical and molecular biology, then for sequence alignment two sequences are homologous if they have descended through a chain of replication from a common precursor molecule, and their residues are also homologous if they have, in turn, descended through a chain of replication from a common precursor set of residues. If a multiple sequence alignment is to represent homology relationships, then all of the nucleotides in any column of the alignment should be homologous, or at least be hypothesized as homologous. Homology is not the only possible criterion for aligning nucleotides, but it is the one that I am addressing here: homology is the relationship among parts of organisms that provides evidence for common ancestry (Brower and de Pinna 2012).

Sequence alignment is one of the core techniques in bioinformatics (Wallace et al. 2005; Edgar and Batzoglou 2006; Kumar and Filipski 2007; Notredame 2007; Pei 2008; Kemena and Notredame 2009). Indeed, some of the most-cited papers in biology describe the most commonly used alignment programs: BLAST for pairwise alignment (papers ranked 12th and 14th in the Science Citation Index) and Clustal for multiple alignment (ranked 10th and 28th) (van Noorden et al. 2014). Bioinformatics lies at the junction of mathematics, computing and biology. The computer programs implement mathematical algorithms in a usable and efficient way, and the algorithms define a procedure for optimizing some objective function. The objective function will be an equation (or set of equations) that mathematically defines some biological notion, so that optimizing the function with respect to any given data will yield a biologically relevant answer. This nexus defines the importance of bioinformatics in modern biological science.

The catch for sequence alignment is that there is no known objective function for identifying homology, and so the bioinformatics nexus breaks down. Homology relations are defined by unique historical events (Donoghue 1992; Brigandt 2003), which by their very nature are unobservable: homology exists independently of our ability to recognize it. Comparative biology is thus based on studying the features of contemporary organisms, on the grounds that they will contain traces of their historical ancestry, from which homology relations might be extracted, however imperfectly. It is, however, very difficult to get any informatics into this biology.

The mathematical argument for current computerized alignment practices is basically this:

$$\text{similarity} = \text{homology} + \text{analogy}$$

where homology is defined as similarity due to historical relationships by descent, and analogy is similarity due to all other causes (e.g. common function). Moreover,

if analogy → 0

then similarity → homology

This assumption opens up the door to detecting homology solely as similarity, which can be measured in many different ways based on many different features: nucleotide, amino acid, codon, genome, chromosome, phenotype, etc. Statistical analysis of similarity can then be taken to reflect evolutionary (homologous) relationships (Margoliash 1963). This means that molecular evidence currently makes little use of the long-standing distinction between analogy and homology.

For multiple sequence alignment, the most common computer implementations of this approach have a two-step algorithm: (i) maximize the pairwise similarity of the sequences, often using dynamic programming (which is an exact algorithm); and then (ii) maximize the sum-of-pairs similarity for multiple sequences, often using progressive alignment (which is a heuristic algorithm). The objective function being maximized is based on nucleotide identity, or defined with respect to a substitution matrix (a log-odds scoring scheme), and a penalty for introducing gaps into the sequences (representing "indels"). This approach has generally been called "optimal alignment" in the bioinformatics literature. However, it has long been known that even an optimal pairwise sequence alignment is not necessarily the homologous alignment (Fitch and Smith 1983).

The basic limitation of this approach is that analogy becomes detectably non-zero as evolutionary distance increases (Simmons and Freudenstein 2003). If we want >90% success of the alignment process, then we need >80% nucleotide identity (Morrison 2006) or <0.5 substitutions per site (Fig. 1). This means that optimal alignment fails in situations where many (most?) of the interesting biological questions occur (Morrison 2006; Phillips 2006). It fails particularly where there are closely adjacent but independent insertions / deletions (Golubchik et al. 2007). Knowledge of the ancestral sequences helps, but not completely (Morrison 2009a).

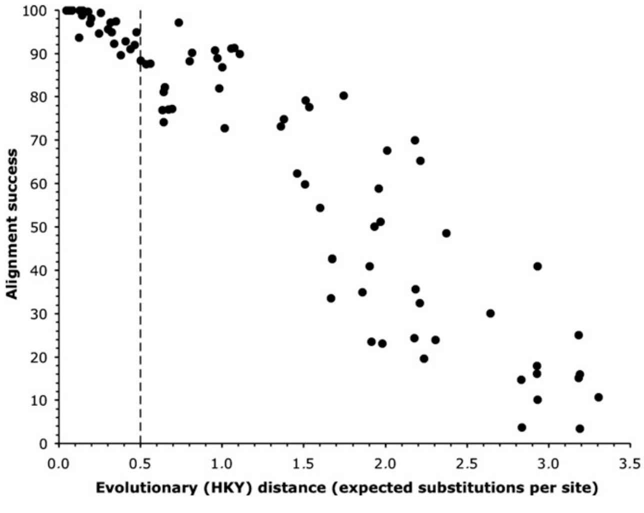

FIG. 1. An example of the relationship between genetic distance and alignment correctness, showing the rapid decrease in accuracy as the distance between sequences increases. Each point represents a single nucleotide alignment generated using the Rose simulation program (see Morrison 2006 for a description of the parameters used). The subsequent multiple alignments were produced with the ClustalW 1.83 program, with default settings. The score shows the average percentage of alignment positions that were correctly aligned by comparison with the reference alignment.

One solution to this problem has been to abandon the idea that residue homology is necessarily the main criterion for sequence alignment, as described below. That is, "accurate alignment" and "homolog recognition" may be quite different things. Indeed, this is true for many of the purposes to which alignments are applied. Only in phylogenetics has homology remained the goal, and most practitioners try to achieve this manually rather than relying solely on computer programs (Morrison 2009b).

USES OF MULTIPLE ALIGNMENTS

Homology is a hierarchical concept (Donoghue 1992), and there are many different components to it related to different levels of biological organization. Morrison et al. (2015) suggest that these levels of molecular complexity each have their own theoretical and practical goals, so that homology between features can simultaneously be present at one level of the hierarchy but absent at others: evolutionary homology, character-state homology, character homology, locus homology, structural homology, genic homology, developmental homology, and taxic homology.

Homology at one level in the hierarchy does not necessarily provide information about homology at any other level in the hierarchy. Formally, a statement of historical identity at one level of the hierarchy requires the historical identity of features at all more inclusive levels but not less inclusive ones. Thus, nucleotide homology is not necessary for the study of macromolecule (gene) homology, because molecules are "above" nucleotides in the hierarchy (they are more inclusive). That is, knowledge of nucleotide homology is not necessary for the study of macromolecule homology, even if it may be useful in many cases.

Recognizing this has opened up the way for specialist approaches to multiple alignment, each with its own separate objective and algorithms. Multiple sequence alignment has thus fragmented into at least four different strands (Morrison 2006): structure prediction; database searching; sequence comparison; and phylogenetic analysis. While this fragmentation of purpose has been recognized in the bioinformatics literature (e.g. Friedrich et al. 2009), the fact that these purposes may require distinct alignments is rarely recognized (Morrison 2006).

Only phylogenetic analysis necessarily requires nucleotide homology, as taxic homology is required for a study of organismal relationships. Thus, while alignment and homology are usually mentioned in the same breath, in practice this connection has been irrelevant for almost all automated sequence alignment programs. Instead, we have optimal alignments, whose adjectives are (based on a casual look through the literature): good, high-quality, accurate (very popular), reliable (also popular), realistic, improved, etc. None of these appear to have formal mathematical definitions that would allow an optimization algorithm, although "accuracy" is presumably being used in the statistical sense representing closeness to the true value. However, the "true value" for real data in this case is alleged to be homology, which we cannot measure!

Structure prediction attempts to deduce the secondary and tertiary structure of the gene product from the gene sequence. For this to be successful, we need to align structurally equivalent nucleotides, irrespective of whether they are homologous or not. That is, spatially related residues in the gene product may have had their structural roles shifted during history, and bioinformatics has developed algorithms for predicting structure based on (current) spatial relations rather than on (historical) homology relations (Shapiro et al. 2007; Zhang 2008).

Database searching attempts to maximize the distinction between homologous and non-homologous sequences, thus allowing the detection of homologous genes in a large collection of sequences. This can be done without necessarily identifying each and every homologous nucleotide in those sequences (Griffiths-Jones and Bateman 2002), which is especially important when sequence identity is small. Bioinformatics has recently developed algorithms for combining multiple sequences based on positional probabilities, such as conserved motifs, profile alignments and Markov models, which markedly improve database searching (Fuellen 2008; Li et al. 2011). The use of "intermediate" sequences is also common, which do not need to be homologous sequences but merely sequences showing large identity to the query sequence (see below).

Sequence comparisons juxtapose residues representing conserved sequence features (e.g. functional assignments, catalytic residues, interaction networks, mutation sites). All of the commonly used multiple-alignment programs were originally developed for sequence comparison; but many specialist programs are now being developed for particular subsets of sequence comparison, such as motif recognition and binding sites. For this to be successful we need to align functionally related residues; and functional roles may shift between residues during history, thus destroying the connection to evolutionary homology. Bioinformatics has developed algorithms for combining different alignment types (e.g. local, global), consistency and probabilistic measures have replaced sum-of-pairs as the optimality criterion, and iterative refinement has been shown to be fast and effective. This has been reviewed many times recently (Thompson and Poch 2005; Wallace et al. 2005; Edgar and Batzoglou 2006; Morrison 2006; Phillips 2006; Notredame 2007; Pei 2008; Do and Katoh 2009; Kemena and Notredame 2009; Anisimova et al. 2010; Löytynoja 2012; Ortuño et al. 2013a; Warnow 2013).

Phylogenetic analysis produces hypotheses of evolutionary relationships among the sequences. For this, nucleotides should be aligned only if they have descended from a common ancestral nucleotide. Reliable phylogenetic estimation requires the comparison of homologous characters, as only these comparisons can be justified as retaining the hierarchical signal of descent with modification. Thus, nucleotide homology is the essential component in the distinction of phylogenetic alignment from alignment for the other three purposes (Morrison 2006; Phillips 2006); mathematically, homology is both necessary and sufficient for phylogenetic analysis.

Unfortunately, bioinformatics has not developed much in the way of algorithms for phylogenetic alignment. There are basically three practical strategies used by phylogeneticists: (1) try to use the computer programs developed for one of the other three purposes, particularly those for sequence comparison (e.g. Clustal, MAFFT, ProbCons); (2) combine alignment and tree-building, via statistical alignment or direct optimization (discussed below); and (3) manually intervene in the alignment process, in an attempt to correct obvious "mistakes" with respect to likely homology.

A survey of the literature shows that strategy 1 is the most common strategy in practice, but that more than one-half of evolutionary biologists intervene manually in their sequence alignments (strategy 3), and more than three-quarters of phylogeneticists do so (Morrison 2009b). Apparently, while personal judgment may not be perfect, at least it can be consciously based on homology as a concept, which computerized algorithms cannot (yet).

Manual alignments are certainly tedious to produce, and they may also be unrepeatable unless there is a clearly stated objective and a protocol. On the other hand, automatic alignments are speedy and convenient, but they are unlikely to fully reflect homology, at least under realistic conditions of length-variable sequences. The distinction between so-called "automatic" and "manual" alignment procedures is sometimes presented as important (Giribet et al. 2002; Anisimova et al. 2010); however, this is a red herring (Kjer et al. 2007; Morrison 2009b). Any automatic procedure can be done manually (given enough time) and any manual procedure can be automated (given an algorithm). The distinction is merely one of convenience: the main practical problem with manual alignment is that it is inefficient (see Kawrykow et al. 2012), rather than that it might be subject to investigator bias. Bias, for example, is also manifested in computerized methods, both by arbitrary choice of program parameters (Morrison and Ellis 1997) and by the choice of guide tree. This is discussed further by Morrison et al. (2015).

Alignment and Homology Criteria

The problem here seems to be the lack of a clear focus on what homology means in biology, and how that meaning can be transferred to molecular sequences in a way that allows development of an objective function (or at least a heuristic algorithm). Perhaps this is because the word "homology" can have several somewhat different meanings (Donoghue 1992; Brigandt 2003).

Homology was originally defined (by Richard Owen) without reference to evolutionary history, but with the Darwinian revolution the term came to have an explicitly evolutionary meaning based on common ancestry. This interpretation of homology can be expressed by the term "homogeny" (as originally proposed by Ray Lankester), and it is this sense that it is important for phylogenetic analysis. Other possible interpretations of homology as "essential sameness" include similarity that may be a consequence of functional or structural relationships; and this explains the common claim that multiple sequence alignments can validly align residues based on structural/functional equivalence as the primary criterion (Thompson and Poch 2005; Edgar and Batzoglou 2006; Pei 2008; Do and Katoh 2009). For this reason, in molecular biology "homology" is often treated as a synonym of "similarity" (Marabotti and Facchiano 2009), usually measured as percent identity.

For evolutionary purposes, we need to focus on homology based on common ancestry (Donoghue 1992; Freudenstein 2005). Traditionally (i.e. for phenotypic data), characters have been first proposed to be homologous using the criteria of similarity and conjunction (together called primary homology), and then tested with the criterion of congruence (secondary homology) (de Pinna 1991; Brower and Schawaroch 1996; Hawkins et al. 1997). Clearly, we need to understand these three concepts and the processes involved, if we are to apply them to molecular data. Unfortunately, previous considerations of molecular homology have focused on gene homology rather than on nucleotide homology (Fitch 1970; Patterson 1988; Hillis 1994; Brigandt 2003; Freudenstein 2005; Haggerty et al. 2014).

Homologous features should pass the three tests of similarity, conjunction and congruence (Patterson 1982); and these can thus be used as criteria for detecting potential homologies (Table 1). Similarity refers to any apparent sameness of the features, and it is the usual criterion for recognizing the possibility of homology. Conjunction refers to the fact that truly homologous characters cannot co-exist in an organism at the same time, when one is referring to taxic homology. Congruence refers to co-occurring distributions of character states on a phylogenetic tree (i.e. multiple characters have states with consistent distributions).

These criteria may contradict each other when assessing homology (see below), and so congruence is often seen as the decisive test of homology (Patterson 1982) because it is the only one directly tied to ancestry. However, all we have is an estimate of ancestry, which may be wildly inaccurate, and so this decisiveness may be misconstrued.

When dealing with nucleotide alignment, these original "tests" for phenotype characters instead become optimization criteria when applied to genotype. For example, the test of similarity as a criterion for recognizing homology becomes a criterion for a mathematical algorithm whose goal is an optimal alignment, which is then treated as the best hypothesis of homology. For phenotype, hypotheses of homology are proposed and then tested, but for genotype the test criterion is also the proposal criterion.

So, I will now try to conceptually relate each of these "tests" as potential optimization criteria for the alignment of nucleotide sequences. There is a simple connection between the homology criteria and the characteristics of nucleotide sequences, although this appears never to have been discussed before (which is surprising, given the voluminous literature on homology).

Similarity—Similarity obviously plays a major role in recognizing potential homologies, and it can be considered to have several components (Remane 1952): Compositional, Topographical, Ontogenetic, and possibly Functional. These criteria need to be related directly to nucleotide alignment, if we are to reliably detect homology; so I will discuss them in turn.

COMPOSITIONAL SIMILARITY—Compositional similarity refers to apparent likeness or resemblance. For nucleotide sequences this is not based on likeness of the units of comparison themselves (the four types of nucleotides: A, C, G, T), but is based on likeness of the character states in each character. Compositional similarity is usually measured as percent identity, based on the substitution scoring scheme in use and the penalty given to indels. It is the sole basis of almost all of the computer programs that produce "optimal" alignments, irrespective of their degree of algorithmic sophistication. It has even been suggested that all alignments *should* be possible ab initio, using only the (compositional) information in the sequences themselves (Kemena and Notredame 2009), and to this end much of the algorithmic sophistication of recent programs has been directed at compositional similarity alone (sometimes then called positional homology). However, for nucleotides (of which there are only four) compositional similarity will be random at 25% sequence identity.

However, compositional similarity is not on its own a viable criterion for homology. That is, the optimal alignment (= maximum compositional similarity) is not necessarily the correct alignment, irrespective of how "correct" is defined biologically (Morrison 2006). Therefore, optimal alignment has probably been the biggest hindrance to the development of an automated homology method, as it has distracted both the theory and practice away from the main biological issues into computational ones that have shown little promise for dealing with homology assessment.

Moreover, compositional similarity as a criterion for homology is implemented inappropriately in many computer programs, because they assume that substitutions and indels occur at random. This has rarely been explicitly pointed out (Redelings and Suchard 2007; Lunter et al. 2008) or its importance stressed. Sequence variation occurs distinctly non-randomly, due to the molecular mechanisms causing sequence variation, and this has direct consequences not only for alignment but also for assessments of alignment accuracy and simulation studies of alignment algorithms. Sequence variation is anything but random (even locally; Terekhanova et al. 2013), and the non-randomness contains valuable information about homologies. For this reason, I believe that we have not yet developed a simulation procedure that comes even close to being realistic for evaluating the success of multiple alignment algorithms in terms of homology.

These issues are compounded by the fact that all length-invariable differences among the sequences are modeled as substitutions, and all length-variable differences are modeled as indels. These are very simplistic models, given that sequence variation is caused by a range of molecular mechanisms, including (Table 2): duplications, notably tandem repeats (copying to an immediately adjacent position) and inverted repeats (reverse-complementing the copy); substitutions; inversions; translocations and transpositions; deletions; and insertions. Some of these processes create length variation in the sequences, while others do not, and some

TABLE 1. Tests used to evaluate potential homologies among nucleotide sequences

Homology test	Criterion for nucleotides	Measurement form
Similarity		
Compositional	Apparent likeness or resemblance between sequences	% identity
Topographical	Spatial relationship to other characters in the same sequence	Second- and third-order structure
Functional	Functional relationship to other characters in the same sequence	Annotated function
Ontogenetic	Variation arising from the same molecular mechanism between sequences	Inferred molecular mechanism
Conjunction	Absence of multiple copies in the same sequence	Number of copies
Congruence	Agreement with other postulated homologies in the same sequences	Synapomorphy

Table 2. Mechanisms causing sequence variation, and their consequences for the homology tests. * Following the terminology of Patterson (1988).

Molecular mechanism	Within-gene relationship*	Test result
Substitution	Orthology	Homology
Inversion (replacement of a subsequence by its reverse-complement)	Orthology	Homology
Translocation (removal of a subsequence and its insertion at another location)	Xenology	Fails congruence
Transposition (exchange of subsequences between locations)	Xenology	Fails congruence
Duplication (copying a subsequence, e.g. tandem repeat, inverted repeat)	Paralogy	Fails conjunction
Insertion (addition of a novel subsequence)	Complement	Fails similarity
Deletion (removal of an existing subsequence)	Complement	Fails similarity

of them violate the usual assumption of alignment algorithms that all of the sequences are co-linear. Of these processes, tandem repeats are probably the most common cause of sequence variation (Huntley and Clark 2007; Messer and Arndt 2007); and my own experience is that they are the phenomena most often mislead alignment programs that are based on compositional similarity.

Topographical Similarity—Topographical similarity (or topological correspondence) uses relationships within the same sequence to identify possible homologies, whereas compositional similarity relies solely on relationships between sequences. For example (attributable to Richard Owen), anatomical assessment of the homology of the bones of vertebrate limbs is based on the number and spatial relationship of the bones within each of the limbs (topographical similarity), not on their size or shape (compositional similarity).

Topographical similarity of nucleotide sequences is based on the second-order (planar) and third-order (three-dimensional) structure of the encoded gene product. For example, for RNA-coding sequences (and non-coding sequences such as group I & II introns, and transcribed spacers), the most important second-order structures are the double-stranded stems, while codons make up the most important second-order structural feature of protein-coding sequences (Morrison 2009a).

These structures place non-random biological constraints on the macromolecule, and different regions of the sequence have different functional constraints. Where stem-pairing or codons enforce constraints, compositional similarity will not necessarily be maintained, and only topographical similarity will be an effective criterion for recognizing homology. For example, Illergård et al. (2009) noted that protein structure diverges three to ten times more slowly than the sequence itself.

Indeed, compositional similarity only works in highly conserved regions; and in the so-called regions of ambiguous alignment or regions of expansion and contraction (Gillespie 2004), even topographical similarity will fail, because here topology is not conserved either.

There are currently no general computer programs based directly on topographical similarity as a criterion for multiple alignment. There are, however, programs that will do this indirectly, for example, by translating nucleotide sequences to amino acid sequences and then using compositional or topographical similarity for alignment of the amino acid sequences (Pei et al. 2008; Herman et al. 2014); and there are criteria by which the success of these alignments can be judged (Higgins and Taylor 2000; Lebrun et al. 2006). There are also programs that will attempt to align RNA-coding sequences based on their inferred secondary structure (Tabei et al. 2008; Wilm et al. 2008; Sahraeian and Yoo 2011; DeBlasio et al. 2012) as well as tertiary structure (Kemena et al. 2013), but it is clear that they are currently inadequate for the data sets used in phylogenetics (Letsch et al. 2010); and so this criterion is mostly implemented manually. Several published papers have provided suitable objective criteria for manually creating these RNA alignments (Kjer et al. 1994, 2009; Kjer 1995; Gillespie 2004).

Functional Similarity—Functional similarity is sometimes mentioned as a separate criterion, but it is directly analogous to topographical similarity and could be subsumed within it. Structure and function are usually intimately related, although they do not necessarily exist in a simple 1:1 relationship to each other. Indeed, neither structure nor function is necessarily well-defined in a set of sequences, as single nucleotides may have multiple functions (or be involved in multiple structural features) and multiple nucleotides may share a single function. Since structure and function may change through evolutionary history (Wray and Abouheif 1998), these two criteria may contradict inferences from each other and from the other similarity criteria. Indeed, functional similarity might be quantified differently by biochemists, biophysicists, immunologists, structural biologists, etc. Annotations of function come from experimental data, the access to which is increasing rapidly with the development of new technologies; and so this may be important for future alignment methods. There are many specialist programs, for example for motif finding (e.g. Das and Dai 2007), which are particularly prone to inversions and transpositions (Gordân et al. 2010).

Ontogenetic Similarity—Ontogenetic similarity refers to the use of developmental processes and timing as criteria for recognizing homology. As a well-known example in botany, morphological assessment of the homology of the phyllodes of *Acacia* and the petioles of bipinnate leaves is based on observations of the transitional forms during their development (ontogenetic similarity), not on their size or shape (compositional similarity).

Ontogenetic similarity of nucleotide sequences is based on identifying the known molecular processes that cause sequence variation, as discussed above. That is, the multiple alignment would represent a set of scenarios for how the set of sequences developed, in terms of some combination of repeats, inversions, insertions, etc. These are explicitly inferred comparisons within each sequence (e.g. for repeats) and between sequences (e.g. for inversions).

There are currently no computer programs for global multiple alignment of nucleotides based directly on ontogenetic similarity. The issue with algorithms that do not explicitly account for specific mechanisms is that the simple indel/substitution model mis-weights the events; for example, a single 4-base inversion would be counted as four substitutions. Work has been done for single sequences involving repeats (see Morrison 2006; Leclercq et al. 2007) and inverted repeats (Warburton et al. 2004; Gupta et al. 2006; Sreeskandarajan et al. 2014); and for pairwise alignments involving duplications (Bertrand and Gascuel 2005; Sammeth and Stoye 2006; Freschi and Bogliolo 2012), inversions (Schöniger and Waterman 1992; Chen et al. 2004; Pereira do Lago et al. 2005; Vellozo et al. 2006), rearrangements (Chu et al. 2009), and their combination (Hsu and Cull 2001; Cull and Hsu 2003; Ledergerber and Dessimoz 2008). There are also programs based on multiple local alignments of deleted, repeated or rearranged sequence blocks but not individual nucleotides (Darling et al. 2004; Phuong et al. 2006; Blanco et al. 2007).

So, to date this criterion has been implemented manually for DNA sequences. Several published papers have provided suitable objective criteria for creating the alignments (Golenberg et al. 1993; Kelchner and Clark 1997; Hoot and Douglas 1998; Graham et al. 2000; Borsch et al. 2003; Löhne and Borsch 2005; Benavides et al. 2007). Alternatively, the use of profiles of so-called intermediate sequences as templates for the pairwise alignment step (in programs such as Praline, Promals and PSI-Coffee) can be seen as a heuristic attempt to incorporate some of the information from ontogeny into multiple alignment (Simossis and Heringa 2005; Pei and Grishin 2007; Kemena and Notredame 2009).

Ontogeny-based alignments have the advantage of making clear that homology refers to the relationship between parts of organisms that have resulted from the same heritable transformation event (Morrison 2009a). These events should be used in creating the alignment rather than being a posteriori deductions from it. Hypotheses about these evolutionary events should be plausible and parsimonious; and it should be possible to list these events as a supplement to the tabular alignment presentation (empirical examples are shown by Löhne and Borsch 2005; Müller and Borsch 2005; Borsch et al. 2007).

Conjunction—Conjunction (or coexistence) as a criterion or test of taxic homology simply states that homologous features cannot have multiple copies within the same organism. If the features were to be repeated, then there would be no way to determine which of them is homologous to the single copy in another organism. This criterion is not often applied (Patterson 1982, 1988), except that it is considered to be an important component in developmental biology, where homology within organisms is as important as homology between them (Brigandt 2003). It has clear implications for molecular sequences (Holland 1999), where gene duplications and sequence repeats are common (see the next section). Furthermore, gene products are expressed at levels of more than one copy per cell, and there are many greatly repetitive genes (such as rRNA and histones), so a rigid interpretation of conjunction is not practical (States and Boguski 1991).

Duplications and other within-gene repeats involve serial homology, which then becomes an important conceptual issue for alignment. In a multiple alignment an explicit decision is made to align certain nucleotides and not others, and a decision therefore needs to be made about whether to align duplications and repeats, and if so how to do so in practice. For example, the distinction between orthologous and paralogous genes needs to be explicitly addressed, as mixing them causes confusion in phylogenetic analyses (de Pinna 1991); and the same issue applies at a finer scale, with tandem repeats, for example. Presence or absence of repeats provides information about between-organism homology, but the actual number of repeats is often related only to within-organism serial homology.

Congruence—Congruence as a test of homology recognizes that congruent patterns among multiple postulated homologies provide strong evidence that the homologies are correct. For example, in anatomy, bird wings and bat wings do not pass the congruence test because they do not form a synapomorphy on a phylogenetic tree produced by other anatomical data. That is, there are apparently two separate origins of vertebrate wings.

There are currently two computerized approaches that use congruence as a criterion for nucleotide alignment, both of which operate by combining alignment and tree-building into a single procedure: statistical alignment (Lunter et al. 2005; Westesson et al. 2012a) and direct optimization (Phillips et al. 2000; Wheeler et al. 2015). In essence, these approaches build trees under models that include indels in addition to substitutions.

Statistical alignment adopts a probabilistic approach to alignment and tree-building. Explicit models of sequence evolution are constructed in a likelihood context, incorporating both substitutions and indels as explicit evolutionary events (Metzler and Fleissner 2009). Some criterion is then used to optimize the parameters in relation to the model, such as maximizing either the likelihood or the Bayesian posterior probability. Thus, statistical alignment tries to combine compositional similarity and congruence as criteria for homology.

However, this approach maximizes dependence among the characters, by confounding compositional similarity and congruence rather than treating them as independent criteria (Simmons 2004; Morrison 2009a: Simmons et al. 2010). A duality between alignment and tree building has long been recognized (Sankoff et al. 1973), although not necessarily in the context of homology assessment, but this does not mean that alignment and tree building must be inextricably confounded (Mindell 1991). We need to evaluate the strength of similarity (primary homology) as a criterion in different parts of the alignment, as well as the strength of congruence (secondary homology), because homology is actually an inference problem rather than an optimization problem (Morrison 2009a). Furthermore, current implementations of statistical alignment use rather unsophisticated models (Morrison 2009a), where the indel model is overly simplistic to the extent of being misleading.

Direct optimization optimizes ancestral sequences on a tree while treating gaps as a character state rather than as missing data. The correct alignment is seen to be the one that produces the minimum-cost phylogenetic tree (evaluated via parsimony or likelihood), where all of the cost parameters (substitution costs, gap penalties, sequence weights, etc.) are specified concurrently for both the alignment and the tree. Thus, the alignment is simply the tabular version of the homologies while the tree is the graphical version. This approach also maximizes dependence among the characters (Simmons 2004; Simmons et al. 2010), and so suffers similar problems to statistical alignment.

Furthermore, direct optimization makes congruence the sole arbiter of homology, treating synapomorphy and homology as being synonymous (so-called taxic homology). However, two wrongs don't make a right, and so mere congruence of characters alone cannot determine homology

(Nixon and Carpenter 2012; Assis 2013). While homology implies synapomorphy, apparent synapomorphy does not necessarily imply homology. One needs an independent causal basis for the hypotheses of homology, involving theories of inheritance and development (Morgan and Kelchner 2010), which is provided by the similarity and conjunction criteria. Thus, direct optimization is unlikely to be a successful procedure for detecting nucleotide homologies, either in theory (Simmons 2004; Rieppel 2007; Simmons et al. 2008; Morrison 2009a; Yoshizawa 2010; Morgan and Kelchner 2010) or in practice (Cognato and Vogler 2001; Huttunen and Ignatov 2004; Petersen et al. 2004; Aagesen et al. 2005; Creer et al. 2006).

Differences in Alignment Criteria—The criteria of compositional, topographical and ontogenetic similarity are reasonably independent, and so alignments based on them do not necessarily produce the same set of hypotheses of nucleotide homology (Remane 1952). That is, the evidence for homology is not necessarily unambiguous. Here I illustrate empirically what can happen if each criterion is applied on its own.

Figure 2a shows the alignment of six sequences based on compositional similarity, while Fig. 2b shows the same data aligned based on topographical similarity (showing that the sequences form a stem-loop rRNA structure). Nucleotides that are involved in the same stem-pairs have been aligned, while the loop nucleotides are unaligned. The stem lengths differ between the sequences, so that not all of the sequences are involved in each alignment column. Figure 2c shows the same data aligned based on ontogenetic similarity. Version (i) of this alignment shows the hypothesized ancestral alignment, while (ii) and (iii) indicate that the subsequent sequence differences originate as two sets of tandem repeats. The first one (ii) involves an AAAT motif (boxed) in sequences5 + 6 and the second one (iii) involves an AAA in sequences3 + 4. The three alignments (a), (b) and (c(iii)) are clearly very different from each other, representing conflicting hypotheses of homology — only the first and last five alignment columns are shared in common.

A. Compositional similarity

```
sequence1  CTTAAT    TCATTTGAG
sequence2  CTTAAA    TAATTTGAG
sequence3  CCTAAA AAATAATTTGAG
sequence4  CCTAAA AAATAAATTGAG
sequence5  CTTAAATAAATAATTTGAG
sequence6  CTTAAATAAATAATTTGAG
```

B. Topographical similarity

```
sequence1  |CTTAA  | [TTCAT   ] |  TTGAG|
sequence2  |CTTAAA | [TAA     ] | TTTGAG|
sequence3  |CCTAAA | [AAATAA  ] | TTTGAG|
sequence4  |CCTAA  | [AAAATAAA] |  TTGAG|
sequence5  |CTTAAAT| [AAATA   ] |ATTTGAG|
sequence6  |CTTAAAT| [AAATA   ] |ATTTGAG|
```

C. Ontogenetic similarity

```
           (i)                (ii)                   (iii)
sequence1  CTTAATTCATTTGAG    CTTAATT    CATTTGAG    CTTAAT   T    CATTTGAG
sequence2  CTTAAATAATTTGAG    CTTAAAT    AATTTGAG    CTTAAA   T    AATTTGAG
sequence3  CCTAAATAATTTGAG    CCTAAAT    AATTTGAG    CCTAAAAAAT    AATTTGAG
sequence4  CCTAAATAAATTGAG    CCTAAAT    AAATTGAG    CCTAAAAAAT    AAATTGAG
sequence5  CTTAAATAATTTGAG    CTTAAATAAATAATTTGAG    CTTAAA   TAAATAATTTGAG
sequence6  CTTAAATAATTTGAG    CTTAAATAAATAATTTGAG    CTTAAA   TAAATAATTTGAG
```

FIG. 2. Alignment of six nucleotide sequences, which are part of the *trn*L intron in some bryophytes. A. Alignment optimization based on compositional similarity. B. Alignment optimization based on topographical similarity. C. Alignment optimization based on ontogenetic similarity. In B the sequences are shown to form a stem-loop rRNA structure, with || used to delimit the two (paired) halves of the stem and [] to delimit the loop region. In C there are three alternative versions of the alignment, labeled (i), (ii) and (iii); the duplicated motif is boxed.

A. Topographical similarity

```
amino acid      a  b  c  d  e  f  g  h  i  j
codon position 123123123123123123123123123123
sequence1      TTGACTCCCTTCCCA      TATCTTGCA
sequence2      TCGACCCCCCTCCCATTCCCATATCTTGCA
sequence3      TCCATTCCCCTTAGA      TATCTTTCA
sequence4      TTCACTCCCCTTTTA      TATCTTGTA
```

B. Ontogenetic similarity

```
sequence1      TTGACTCCCTTCCCAT      ATCTTGCA
sequence2      TCGACCCCCCTCCCATTCCCATATCTTGCA
sequence3      TCCATTCCCCTTAGAT      ATCTTTCA
sequence4      TTCACTCCCCTTTTAT      ATCTTGTA
```

FIG. 3. Alignment of four nucleotide sequences, which are part of the *cem*A gene in some plants. A. Alignment optimization based on topographical similarity. B. Alignment optimization based on ontogenetic similarity. In (a) the amino acid positions are indicated by letters and the codon positions by numbers. In B the duplicated motif is boxed.

Figure 3a shows the alignment of four sequences based on topographical similarity. The amino acid positions are indicated by letters and the codon positions by numbers. The 6-base insertion in sequence2 is placed between two amino acids, thus preserving the reading frame, so that amino acids f+g are coded between e+h. Figure 3b shows the same data aligned based on ontogenetic similarity. This shows that the insertion originates as a tandem repeat of a TCCAT motif (boxed). This implies that the Ts aligned in column h1 of the topographical alignment should be aligned in column f1 instead. These two alignments thus have only a relatively small difference, but it affects the reading frame and thus any attempt to analyze the data as amino acid sequences.

It should also be noted that a convention has been used in both figures, where the tandem repeats have been left aligned. That is, the sequences without the repeat have been aligned to the left-most copy of the repeats in the other sequences.

In a similar vein, Blackburne and Whelan (2013) have noted that alignments resulting from those computer programs based on topographical similarity (in this case amino acid sequences) are all rather similar to each other, but are distinctly different from those arising from the programs based on congruence; and Westesson et al. (2012b) have noted the same difference. Also, Kemena et al. (2013) noted that for RNA-coding sequences where <60% of the nucleotides form basepairs compositional similarity often outperforms topographical similarity as a criterion. Furthermore. Simmons et al. (2002) noted that congruence is much more closely related to compositional similarity for nucleotides than for amino acids.

FAILURE TO PASS THE HOMOLOGY TESTS

True homologies (i.e. homogeny) will pass all three of the tests and their components. However, patterns of sequence variation do not necessarily pass all of these tests. So, we need to consider what happens when data fail one or more of the tests (Table 2), and what the consequences are for sequence alignment.

Only substitution and inversion produce truly homologous sequence variation (Table 2), while translocation and transposition fail the congruence test because the moved regions are not related to the adjacent sequence fragments. That is, all length-invariable sequence changes preserve homology only if xenology is treated as “a form of homology”

(Patterson 1988). Duplicated sequence fragments obviously fail the conjunction test, while inserted and deleted fragments have nothing to be "similar" to and so must fail the similarity test. That is, length variation in the sequences fails at least one of the tests, and so it can be considered to generate some form of analogy.

If a multiple alignment is to represent homology, then strictly speaking all of the non-homology should be excluded from the alignment. For example, alignment gaps (insertions and deletions) would be excluded from phylogenetic analyses, contrary to much common practice. However, to me this seems to be rather an extreme viewpoint. For example, if we exclude non-homology then logically we would need to exclude *all* of the versions of any repeats including the original (i.e. the apparent gapped region would be only part of the non-homology).

At one extreme, the response of the direct-optimization approach is to do away with alignments altogether (i.e. trees are the only statements of homology); and at the other extreme naïve acceptance of similarity alignments simply ignores the possibility of non-homology. In between these extremes, some practitioners delete gapped regions as being the most likely locations of non-homology (presumably leaving only those regions well-aligned by compositional similarity), while others manually adjust alignments to rectify the most obvious failures of compositional similarity.

Alternatively, non-homology will be phylogenetically informative if it creates synapomorphies on a tree, and so it seems best to try to include as much of the sequence variation as possible in any multiple alignment, prior to phylogenetic analysis. The important point is to not align (in the same column) features that are not homologous, so that unaligned nucleotides are clearly indicated as non-homologous in the taxic sense. This probably creates more gaps in alignments than practitioners are used to (Higgins et al. 2005).

Explicitly distinguishing homology and non-homology in a multiple alignment can be done by adopting conventions that allow interpretation of the alignment in a useful manner. The basic problem is that the sequential order in which each nucleotide sequence is conventionally presented (5′–3′) is not appropriate for many alignments, especially where inversions, translocations and transpositions are involved. For example, inversions will appear as a clustered set of substitutions if the nucleotides are left in their default order, and the inverted fragment would be better re-inverted for analysis. Translocations and transpositions may appear as indels in the default order, and the moved fragments could be returned to their original location for analysis. Duplications could be dealt with using an arbitrary convention (Morrison 2009a), such as treating the first copy as the "original" to be aligned, or we could minimize the number of columns with mismatches or the number of mismatches per column. Insertions and deletions are currently treated as missing data in most phylogenetic analyses, rather than being deleted.

Conventions have the advantage of making any procedure objective and repeatable, as they avoid variation induced by arbitrary choices. In practice, this tactic is employed by all of the current computer algorithms (Morrison 2009a), although this fact is not always made explicit by their authors. However, many researchers prefer to exclude regions where such conventions are necessary, on the grounds that the consequent evolutionary hypotheses may be arbitrary (e.g. Kelchner 2000; Borsch et al. 2003; Löhne and Borsch 2005).

Clearly, if inferences from the different criteria for homology contradict each other then we will have ambiguous alignments. That is, there will be more than one plausible alternative alignment (Redelings and Suchard 2009). Any automated procedure for multiple sequence alignment must be able to resolve these conflicts in some biologically acceptable manner. Mathematically, this is usually done by assigning weights (or costs) to the different inferences, so that this cost can be optimized for each data set (Kemena and Notredame 2009), or a probability function is used based on a joint probability model (Redelings and Suchard 2009).

Finally, nothing that is said here excludes the possibility that for some, if not many, sequence regions homology will be indeterminable. These are the so-called "unalignable" regions that most practitioners will have encountered at some time (Lee 2001), for which homology assessment remains intractable.

Possible Strategies for Phylogenetic Alignment

Molecular phylogeneticists want repeatable homology assessment for nucleotide sequences, which has never been achieved for phenotypic data (Hawkins 2000). It is, however, worth noting that so-called optimal sequence alignment is simply an application of the algorithm that Jardine (1967) developed for morphological characters (which is rarely used; see Jardine and Jardine 1967; Jardine 1969), but restricted to compositional similarity rather than all forms of similarity. (That algorithm could not handle "migrations or reorientations of parts".)

Nevertheless, by evaluating the nucleotide-sequence alignment strategies against the accepted criteria for homology, I have now made it clear when and why current computerized algorithms do not succeed in consistently detecting homologous nucleotides — each computer program generally relies on a single criterion for homology. For example, optimal similarity alignments fail because they restrict themselves to the criterion of compositional similarity; and direct optimization fails because it restricts itself to the criterion of congruence.

The main exceptions to this generalization are programs such as 4SALE 1.7 (Seibel et al. 2006), Staccato (Shatsky et al. 2006), MO-SAStrE 2.0 (Ortuño et al. 2013b) and MAFFT 7 (Katoh and Standley 2013), where an option is available that heuristically tries to include both compositional and topographical similarity, and StatAlign 2.0, which tries to combine both congruence and topographical similarity of RNA (Arunapuram et al. 2013) and proteins (Herman et al. 2014).

Structure-based and codon-based alignments are often preferred by practitioners because they include topographical similarity as a criterion; and mechanistic alignments are preferred because they include ontogenetic similarity as a criterion. The latter two strategies involve considerable manual intervention and judgment, however, which somewhat defeats the objective of repeatable homology assessment. Instead of being disappointed that alignment programs don't work as well as we would like them to, perhaps we should be surprised that they work as well as they do, given the fact that each of them ignores most of the criteria for detecting homology. Perhaps they *are* closer to a science than to an art, after all.

In bioinformatics, the biology should come first, the mathematics second and then the computing. This has not happened for multiple alignment of nucleotide sequences, if the alignments are intended to represent hypotheses of homology, as algorithmic developments have not been explicitly directed towards the known biological criteria for homology. The mathematical optimality criteria have not been motivated by the processes responsible for generating the sequence data, and so the algorithms are heuristics; and as I have shown here they have mostly ended up strictly adhering to a single criterion, which is inadequate for consistent homology detection (i.e. they are not suited as defining criteria). No one criterion is intrinsically superior to any other criterion in the search for historically unique events, and so we need to evaluate the relative merits of each of them for any one case.

This leaves open the possibility that different criteria might be most suitable under different circumstances, and that it would be the researcher's responsibility to decide which is which. That is, for any given dataset and set of questions to be answered, there might be one "best" approach to sequence alignment. This leaves us with a set of subjective decisions to be made, with as yet imperfect information about how best to make those decisions. We need, instead, some objective and repeatable approach to evaluating the competing criteria for homology.

Obviously, then, what we need is a computerized procedure that will include all of the known criteria for homology assessment. Each criterion on its own is not a complete test of historical identity, and so they must be combined to provide valid hypotheses of homology. However, we will make no further progress towards a mathematical solution to multiple sequence alignment unless we can define the various homology criteria precisely, in terms comprehensible to computationalists. Indeed, we need a non-phylogenetic diagnosis of homology in order to make homology discovery operational (Jardine 1967, 1969; Agnarsson and Coddington 2008).

So, what we have been trying to do is actually ahead of the currently available informatics tools — there are no current mathematical models. I have no explicit solution to this problem, but several theoretical alternatives present themselves, which I will briefly introduce.

Different Strategies—An ontological approach would be to try to reproduce the human approach to homology assessment, which is via homology hypotheses proposed based on similarity and conjunction, which are then tested with congruence (de Pinna 1991), and perhaps followed by iterative revision of the hypotheses. It is not obvious what a direct algorithmic implementation of this approach would look like. However, it is worth pointing out that iterating between an alignment and a tree is a long-standing suggestion (Hogweg and Hesper 1984; Barton and Sternberg 1987; Corpet 1988; Hein 1990; Wheeler and Gladstein 1994; Gotoh 1996; Vingron and von Haeseler 1997; Edgar 2004; Liu et al. 2009), which clearly recognizes the different roles of similarity and congruence in evaluating homologies by advocating successive approximations or reciprocal illumination (Mindell 1991). Furthermore, it should be possible to adapt the idea of Agnarsson and Coddington (2008), which uses parsimony to test the relative merits of the different homology criteria for each empirical case, prior to producing a tree, which would allow iterative revision of the hypotheses.

An alternative, mechanistic approach would be to search the nucleotide sequences for evidence of known molecular processes, and then optimize the combination of these to produce a set of optimal scenarios for the origin of the sequence variation. That is, there would be some cost function that incorporates all of the costs for the individual homology criteria. This is a form of multi-objective optimization (Handl et al. 2007), an approach that has been proposed for pairwise alignments to produce a range of optimal solutions based on two simultaneous criteria (Taneda 2010; Abbasi et al. 2013) or three (Ortuño et al. 2013b). Unfortunately, this could produce a huge combinatorial problem for multiple alignment (Morrison 2009a), since the set of optimal solutions is likely to be large. For example, extending pairwise alignment from a simple model of substitutions + indels, which has $O(n^2)$ time complexity (where n is the sequence length), to a model with non-overlapping transpositions + inversions + tandem duplications requires $O(n^5)$ time complexity (Ledergerber and Dessimoz 2008), let alone the subsequent extension to multiple alignment, which is typically $O(s^2)$ for s sequences (Boyce et al. 2014).

Alternatively, one could evaluate the three (or four) types of similarity independently as the criteria for alignment hypotheses, and represent the hypotheses as a (large) set of local alignments. These local alignments could then be combined into a global alignment using, for example, the algorithms provided in programs such as M-Coffee (Wallace et al. 2006) or DiAlign-TX (Subramanian et al. 2008). This would be followed by adjustment for conjunction using specified conventions (i.e. the potentially ambiguous alignment of duplicates), and then testing for congruence on a tree. There would then be the possibility of subsequent iterative adjustment and re-testing of the alignment. Much work has been done on aligning sequences pairwise based on repeats, inversions and rearrangements (see above), which could produce the set of local alignments. However, combining these into a global alignment seems to be a rather memory-intensive business (Zola et al. 2007), and so this approach may be unsuitable for data sets typical of phylogenetic analysis. Combining the results of different alignment algorithms is a meta-method (Kemena and Notredame 2009), but use of homology as the criterion for choosing the algorithms has not previously been suggested.

One could also adopt the approach of simply producing a large series of global alignments based on different criteria, as done for example by Wang et al. (2012), who created 37 alignments based on different programs, parameters and guide trees, plus manual alignments. This has the advantage of allowing an explicit assessment of the effects of alignment uncertainty, but it does rather smack of giving up on homology detection. A better alternative would be to integrate across the uncertainty represented by the group of multiple alignments (Blackburne and Whelan 2013), as is done by likelihood analyses, thus producing a unique alignment where each position is associated with a likelihood score.

The logical extreme of this variant would be to develop a comprehensive model that includes all of the criteria for homology, and then use Bayesian analysis to integrate across the cloud of alignments with optimal or near-optimal probability. This is currently done for Statistical Alignment using simple substitution/indel models, and has proven to be effective but resource intensive (computer time and memory).

Perhaps the most effective approach would be to start with a curated and trusted seed alignment and then add new sequences to it (Morrison 2006). If nothing else, this avoids "reinventing the wheel" every time a new alignment is required, but more importantly it allows the high quality of the initial alignment to be maintained as the alignment grows in size (Sievers et al. 2013). Most current alignment programs have a feature that allows new sequences to be added to a pre-existing profile alignment, using hidden Markov models, profiles or templates, although some are likely to be more effective than others; and there are a number of databases that can be used as a source of the profile if you don't have one of your own (Morrison 2006). The effectiveness of this strategy has rarely been assessed (Löytynoja et al. 2012), except in relation to its use in online RNA databases (DeSantis et al. 2006; Nawrocki et al. 2009; Gardner et al. 2012; Pruesse et al. 2012), but I suspect that it has enormous potential in practice.

Finally, a purely heuristic alternative would be to use as a suitable starting point an alignment based on compositional similarity, which will work if sequence identity is great enough, and then modify it to represent a scenario of postulated homologies. This approach seems to work better than any currently available fully automated procedure (Morrison 2009a). Furthermore, it is apparently what is currently being done by many practitioners (Morrison 2009b), although its widespread manual implementation somewhat defeats the objective of having repeatable homology assessment. Automating this approach could be done via explicitly specified criteria for adjusting the alignment, as is currently done by programs that iteratively refine pre-existing alignments (Muller et al. 2010; Lyras and Metzler 2014; Roy et al. 2014). Alternately, Blouin et al. (2009) use a support vector machine to transfer manual annotation of one alignment to another.

Interestingly, these approaches bring phylogenetic sequence alignment full circle. In what appears to be the first published study of intra-specific variation using DNA sequences (Kreitman 1983), Martin Kreitman provided a multiple alignment based on explicit recognition of tandem repeats and RNA stem structures (i.e. ontogenetic and topographical similarity). Apparently, this was produced manually — and we have not progressed much since then.

Acknowledgments. Thanks to Scot Kelchner, Mark Simmons and an anonymous referee for helpfully commenting on an earlier version of this manuscript.

Literature Cited

Aagesen, L., G. Petersen, and O. Seberg. 2005. Sequence length variation, indel costs, and congruence in sensitivity analysis. *Cladistics* 21: 15–30.

Abbasi, M., L. Paquete, A. Liefooghe, M. Pinheiro, and P. Matias. 2013. Improvements on bicriteria pairwise sequence alignment: Algorithms and applications. *Bioinformatics* 29: 996–1003.

Agnarsson, I. and J. A. Coddington. 2008. Quantitative tests of primary homology. *Cladistics* 24: 51–61.

Anisimova, M., G. M. Cannarozzi, and D. A. Liberles. 2010. Finding the balance between the mathematical and biological optima in multiple sequence alignment. *Trends in Evolutionary Biology* 2: e7.

Arunapuram, P., I. Edvardsson, M. Golden, J. W. J. Anderson, Á. Novâk, Z. Sükösd, and J. Hein. 2013. StatAlign 2.0: Combining statistical alignment with RNA secondary structure prediction. *Bioinformatics* 29: 654–655.

Assis, L. C. S. 2013. Are homology and synapomorphy the same or different? *Cladistics* 29: 7–9.

Barton, G. J. and M. J. E. Sternberg. 1987. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. *Journal of Molecular Biology* 198: 327–337.

Benavides, E., R. Baum, D. McClellan, and J. W. Sites. 2007. Molecular phylogenetics of the lizard genus *Microlophus* (Squamata: Tropiduridae): Aligning and retrieving signal from nuclear introns. *Systematic Biology* 56: 776–797.

Bertrand, D. and O. Gascuel. 2005. Topological rearrangements and local search method for tandem duplication trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2: 15–28.

Blackburne, B. P. and S. Whelan. 2013. Class of multiple sequence alignment algorithm affects genomic analysis. *Molecular Biology and Evolution* 30: 642–653.

Blanco, E., R. Guigó, and X. Messeguer. 2007. Multiple non-collinear TF-map alignments of promoter regions. *BMC Bioinformatics* 8: 138.

Blouin, C., S. Perry, A. Lavell, E. Susko, and A. J. Roger. 2009. Reproducing the manual annotation of multiple sequence alignments using a SVM classifier. *Bioinformatics* 25: 3093–3098.

Bock, G. R. and G. Cardew, eds. 1999. *Homology*. New York: John Wiley.

Borsch, T., K. W. Hilu, D. Quandt, V. Wilde, C. Neinhuis, and W. Barthlott. 2003. Noncoding plastid *trn*T–*trn*F sequences reveal a well resolved phylogeny of basal angiosperms. *Journal of Evolutionary Biology* 16: 558–576.

Borsch, T., K. W. Hilu, J. H. Wiersema, C. Löhne, W. Barthlott, and V. Wilde. 2007. Phylogeny of *Nymphaea* (Nymphaeaceae): Evidence from substitutions and microstructural changes in the chloroplast *trn*T–*trn*F region. *International Journal of Plant Sciences* 168: 639–671.

Boyce, K., F. Sievers, and D. G. Higgins. 2014. Simple chained guide trees give high-quality protein multiple sequence alignments. *Proceedings of the National Academy of Sciences USA* 111: 10556–10561.

Brigandt, I. 2003. Homology in comparative, molecular, and evolutionary developmental biology: The radiation of a concept. *The Journal of Experimental Zoology* 299B: 9–17.

Brower, A. V. Z. and M. C. C. de Pinna. 2012. Homology and errors. *Cladistics* 28: 529–538.

Brower, A. V. Z. and V. Schawaroch. 1996. Three steps of homology assessment. *Cladistics* 12: 265–272.

Chen, Z.-Z., Y. Gao, G. Lin, R. Niewiadomski, Y. Wang, and J. Wu. 2004. A space-efficient algorithm for sequence alignment with inversions and reversals. *Theoretical Computer Science* 325: 361–372.

Chu, T.-C., T. Liu, D. T. Lee, G. C. Lee, and A. C.-C. Shih. 2009. GR-Aligner: An algorithm for aligning pairwise genomic sequences containing rearrangement events. *Bioinformatics* 25: 2188–2193.

Cognato, A. I. and A. P. Vogler. 2001. Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Systematic Biology* 50: 758–780.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* 16: 10881–10890.

Creer, S., C. E. Pook, A. Malhotra, and R. S. Thorpe. 2006. Optimal intron analyses in the *Trimeresurus* radiation of Asian pitvipers. *Systematic Biology* 55: 57–72.

Cull, P. and T. Hsu. 2003. Recent advances in the walking tree method for biological sequence alignment. *Lecture Notes in Computer Science* 2809: 349–359.

Darling, A. C. E., B. Mau, F. R. Blattner, and N. T. Perna. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394–1403.

Das, M. K. and H.-K. Dai. 2007. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 8: S21.

de Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7: 367–394.

DeSantis, T. Z., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen. 2006. NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research* 34: W394–W399.

DeBlasio, D., J. Bruand, and S. Zhang. 2012. A memory efficient method for structure-based RNA multiple alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9: 1–11.

Do, C. B. and K. Katoh. 2009. Protein multiple sequence alignment. *Methods in Molecular Biology (Clifton, N.J.)* 484: 379–413.

Donoghue, M. J. 1992. Homology. Pp. 170–179 in *Keywords in evolutionary biology*, ed. E. Fox Keller and E. Lloyd. Cambridge: Harvard University Press.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Edgar, R. C. and S. Batzoglou. 2006. Multiple sequence alignment. *Current Opinion in Structural Biology* 16: 368–373.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology* 19: 99–113.

Fitch, W. M. and T. F. Smith. 1983. Optimal sequence alignments. *Proceedings of the National Academy of Sciences USA* 80: 1382–1386.

Freschi, V. and A. Bogliolo. 2012. A lossy compression technique enabling duplication-aware sequence alignment. *Evolutionary Bioinformatics* 8: 171–180.

Freudenstein, J. V. 2005. Characters, states and homology. *Systematic Biology* 54: 965–973.

Friedrich, A., O. Poch, and L. Moulinier. 2009. Strategies for efficient exploration of the informational content of protein multiple alignments. Pp. 271–295 in *Sequence alignment: Methods, models, concepts, and strategies*, ed. M. S. Rosenberg. Berkeley: University of California Press.

Fuellen, G. 2008. Homology and phylogeny and their automated inference. *Naturwissenschaften* 95: 469–481.

Gardner, D. P., W. Xu, D. P. Miranker, S. Ozer, J. J. Cannonne, and R. R. Gutell. 2012. An accurate scalable template-based alignment algorithm. Pp. *237–243* in *Proceedings of the 2012 IEEE international conference on bioinformatics and biomedicine*. Washington: IEEE Computer Society.

Gillespie, J. J. 2004. Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules. *Molecular Phylogenetics and Evolution* 33: 936–943.

Giribet, G., W. C. Wheeler, and J. Muona. 2002. DNA multiple sequence alignments. Pp. 107–114 in *Molecular systematics and evolution: Theory and practice*, eds. R. DeSalle, G. Giribet and W. Wheeler. Basel: Birkhäuser BioSciences.

Golenberg, E. M., M. T. Clegg, M. L. Durbin, J. Doebley, and D. P. Ma. 1993. Evolution of a noncoding region of the chloroplast genome. *Molecular Phylogenetics and Evolution* 2: 52–64.

Golubchik, T., M. J. Wise, S. Eastel, and L. S. Jermiin. 2007. Mind the gaps: Evidence of bias in estimates of multiple sequence alignments. *Molecular Biology and Evolution* 24: 2433–2442.

Gordân, R., L. Narlikar, and A. J. Hartemink. 2010. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Research* 38: e90.

Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *Journal of Molecular Biology* 264: 823–838.

Graham, S. W., P. A. Reeves, A. C. E. Burns, and R. G. Olmstead. 2000. Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *International Journal of Plant Sciences* 161: S83–S96.

Griffiths-Jones, S. and A. Bateman. 2002. The use of structure information to increase alignment accuracy does not aid homologie detection with profile HMMs. *Bioinformatics* 18: 1243–1249.

Gupta, R., A. Mittal, and S. Gupta. 2006. An efficient algorithm to detect palindromes in DNA sequences using periodicity transform. *Signal Processing* 86: 2067–2073.

Haggerty, L. S., P.-A. Jachiet, W. P. Hanage, D. Fitzpatrick, P. Lopez, M. J. O'Connell, D. Pisani, M. Wilkinson, E. Bapteste, and J. O. McInerney. 2014. A pluralistic account of homology: Adapting the models to the data. *Molecular Biology and Evolution* 31: 501–516.

Hall, B. K., ed. 1994. *Homology: The hierarchical basis of comparative biology*. New York: Academic Press.

Handl, J., D. B. Kell, and J. Knowles. 2007. Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4: 279–292.

Hawkins, J. A. 2000. A survey of primary homology assessment: Different botanists perceive and define characters in different ways. Pp 22–53 in *Homology and systematics: Coding characters for phylogenetic analysis*, eds. R. W. Scotland and R. T. Pennington. London: Taylor and Francis.

Hawkins, J. A., C. E. Hughes, and R. W. Scotland. 1997. Primary homology assessment, characters and character states. *Cladistics* 13: 275–283.

Hein, J. 1990. Unified approach to alignment and phylogenies. *Methods in Enzymology* 183: 626–645.

Herman, J. L., C. J. Challis, Á. Novâk, J. Hein, and S. C. Schmidler. 2014. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of sequence and structure. *Molecular Biology and Evolution* 31: 2251–2266.

Higgins, D. G., G. Blackshields, and I. M. Wallace. 2005. Mind the gaps: Progress in progressive alignment. *Proceedings of the National Academy of Sciences USA* 102: 10411–10412.

Higgins, D. G. and W. R. Taylor. 2000. Multiple sequence alignment. *Methods in Molecular Biology (Clifton, N.J.)* 143: 1–18.

Hillis, D. M. 1994. Homology in molecular biology. Pp. 339–368 in *Homology: The hierarchical basis of comparative biology*, ed. B. K. Hall. New York: Academic Press.

Hogweg, P. and B. Hesper. 1984. The alignment of sets of sequences and the construction of phyletic trees: An integrated method. *Journal of Molecular Evolution* 20: 175–186.

Holland, P. W. H. 1999. The effect of gene duplication on homology. Pp. 226–242 in *Homology*, eds. G. R. Bock and G. Cardew. New York: John Wiley.

Hoot, S. B. and A. W. Douglas. 1998. Phylogeny of the Proteaceae based on *atp*B and *atp*B–*rbc*L intergenic spacer region sequences. *Australian Systematic Biology* 11: 301–320.

Hsu, T. and P. Cull. 2001. Gene verification and discovery by walking tree method. *Pacific Symposium on Biocomputing* 6: 287–298.

Huntley, M. A. and A. G. Clark. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Molecular Biology and Evolution* 24: 2598–2609.

Huttunen, S. and M. S. Ignatov. 2004. Phylogeny of the Brachytheciaceae (Bryophyta) based on morphology and sequence level data. *Cladistics* 20: 151–183.

Illergård, K., D. H. Ardell, and A. Elofsson. 2009. Structure is three to ten times more conserved than sequence – a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* 77: 499–508.

Jardine, N. 1967. The concept of homology in biology. *The British Journal for the Philosophy of Science* 18: 125–139.

Jardine, N. 1969. The observational and theoretical components of homology: A study based on the morphology of the dermal skull-roofs of rhipidistian fishes. *Biological Journal of the Linnean Society. Linnean Society of London* 1: 327–361.

Jardine, N. and C. J. Jardine. 1967. Numerical homology. *Nature* 216: 301–302.

Katoh, K. and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kawrykow, A., G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, Phylo players, L. Sarmenta, M. Blanchette, and J. Waldispühl. 2012. *PLoS ONE* 7: e31362. Phylo: A citizen science approach for improving multiple sequence alignment

Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* 87: 482–498.

Kelchner, S. A. and L. G. Clark. 1997. Molecular evolution and phylogenetic utility of the chloroplast *rpl*16 intron in *Chusquea* and the Bambusoideae (Poaceae). *Molecular Phylogenetics and Evolution* 8: 385–397.

Kemena, C., G. Bussotti, E. Capriotti, M. A. Marti-Renom, and C. Notredame. 2013. Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package. *Bioinformatics* 29: 1112–1119.

Kemena, C. and C. Notredame. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25: 2455–2465.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. *Molecular Phylogenetics and Evolution* 4: 314–330.

Kjer, K. M., G. D. Baldridge, and A. M. Fallon. 1994. Mosquito large subunit ribosomal RNA: Simultaneous alignment of primary and secondary structure. *Biochimica et Biophysica Acta* 1217: 147–155.

Kjer, K. M., J. J. Gillespie, and K. A. Ober. 2007. Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. *Systematic Biology* 56: 133–146.

Kjer, K. M., U. Roshan, and J. J. Gillespie. 2009. Structural and evolutionary considerations for multiple sequence alignment of RNA, and the challenges for algorithms that ignore them. Pp. 105–149 in *Sequence alignment: Methods, models, concepts, and strategies*, ed. M. S. Rosenberg. Berkeley: University of California Press.

Kleisner, K. 2007. The formation of the theory of homology in biological sciences. *Acta Biotheoretica* 55: 317–340.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304: 412–417.

Kumar, S. and A. Filipski. 2007. Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Research* 17: 127–135.

Lebrun, E., J. M. Santini, M. Brugna, A.-L. Ducluzeau, S. Ouchane, B. Schoepp-Cothenet, F. Baymann, and W. Nitschke. 2006. The *rieske* protein: A case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction. *Molecular Biology and Evolution* 23: 1180–1191.

Leclercq, S., E. Rivals, and P. Jarne. 2007. Detecting microsatellites within genomes: Significant variation among algorithms. *BMC Bioinformatics* 8: 125.

Ledergerber, C. and C. Dessimoz. 2008. Alignments with non-overlapping moves, inversions and tandem duplications in O(n^4) time. *Journal of Combinatorial Optimization* 16: 263–278.

Lee, M. S. 2001. Unalignable sequences and molecular evolution. *Trends in Ecology & Evolution* 16: 681–685.

Letsch, H. O., P. Kück, R. R. Stocsits, and B. Misof. 2010. The impact of rRNA secondary structure consideration in alignment and tree reconstruction: Simulated data and a case study on the phylogeny of hexapods. *Molecular Biology and Evolution* 27: 2507–2521.

Li, Y., N. Chia, M. Lauria, and R. Bundschuh. 2011. A performance enhanced PSI-BLAST based on hybrid alignment. *Bioinformatics* 27: 31–37.

Liu, K., S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324: 1561–1564.

Löhne, C. and T. Borsch. 2005. Molecular evolution and phylogenetic utility of the *pet*D Group II intron: A case study in basal angiosperms. *Molecular Biology and Evolution* 22: 317–332.

Löytynoja, A. 2012. Alignment methods: Strategies, challenges, benchmarking, and comparative overview. *Methods in Molecular Biology (Clifton, N.J.)* 855: 203–235.

Löytynoja, A., A. J. Vilella, and N. Goldman. 2012. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* 28: 1684–1691.

Lunter, G., A. J. Drummond, I. Miklós, and J. Hein. 2005. Statistical alignment: Recent progress, new applications, and challenges. Pp. 375–405 in *Statistical methods in molecular evolution*, ed. R. Nielsen. New York: Springer.

Lunter, G., A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein. 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Research* 18: 298–309.

Lyras, D. P. and D. Metzler. 2014. ReformAlign: Improved multiple sequence alignments using a profile-based meta-alignment approach. *BMC Bioinformatics* 15: 265.

Marabotti, A. and A. Facchiano. 2009. When it comes to homology, bad habits die hard. *Trends in Biochemical Sciences* 34: 98–99.

Margoliash, E. 1963. Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences USA* 50: 672–679.

Messer, P. W. and P. F. Arndt. 2007. The majority of recent short DNA insertions in the human genome are tandem duplications. *Molecular Biology and Evolution* 24: 1190–1197.

Metzler, D. and R. Fleissner. 2009. Sequence evolution models for simultaneous alignment and phylogeny construction. Pp. 71–93 in *Sequence alignment: Methods, models, concepts, and strategies*, ed. M. S. Rosenberg. Berkeley: University of California Press.

Mindell, D. P. 1991. Similarity *and* congruence as criteria for molecular homology. *Molecular Biology and Evolution* 8: 897–900.

Morgan, M. J. and S. A. Kelchner. 2010. Inference of molecular homology and sequence alignment by direct optimization. *Molecular Phylogenetics and Evolution* 56: 305–311.

Morrison, D. A. 2006. Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany* 19: 479–539.

Morrison, D. A. 2009a. A framework for phylogenetic sequence alignment. *Plant Systematics and Evolution* 282: 127–149.

Morrison, D. A. 2009b. Why would phylogeneticists ignore computerized sequence alignment? *Systematic Biology* 58: 150–158.

Morrison, D. A. and J. T. Ellis. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. *Molecular Biology and Evolution* 14: 428–441.

Morrison, D. A., M. J. Morgan, and S. A. Kelchner. 2015. Molecular homology and multiple sequence alignment: an analysis of concepts and practice. *Australian Systematic Botany* (in press).

Muller, J., C. J. Creevey, J. D. Thompson, D. Arendt, and P. Bork. 2010. AQUA: Automated quality improvement for multiple sequence alignments. *Bioinformatics* 26: 263–265.

Müller, K. and T. Borsch. 2005. Phylogenetics of Amaranthaceae based on *mat*K/*trn*K sequence data — evidence from parsimony, likelihood, and bayesian methods. *Annals of the Missouri Botanical Garden* 92: 66–102.

Nawrocki, E. P., D. L. Kolbe, and S. R. Eddy. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* 25: 1335–1337.

Nixon, K. C. and J. M. Carpenter. 2012. On homology. *Cladistics* 28: 160–169.

Notredame, C. 2007. Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology* 3: e123.

Ortuño, F. M., O. Valenzuela, H. Pomares, F. Rojas, J. P. Florido, J. M. Urquiza, and I. Rojas. 2013a. Predicting the accuracy of multiple sequence alignment algorithms by using computational intelligent techniques. *Nucleic Acids Research* 41: e26.

Ortuño, F. M., O. Valenzuela, F. Rojas, H. Pomares, J. P. Florido, J. M. Urquiza, and I. Rojas. 2013b. Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: Structural information, non-gaps percentage and totally conserved columns. *Bioinformatics* 29: 2112–2121.

Patterson, C. 1982. Morphological characters and homology. Pp. 21–74 in *Problems of phylogenetic reconstruction*, eds. K. A. Joysey and A. E. Friday. London: Academic Press.

Patterson, C. 1988. Homology in classical and molecular biology. *Molecular Biology and Evolution* 5: 603–625.

Pei, J. 2008. Multiple protein sequence alignment. *Current Opinion in Structural Biology* 18: 382–386.

Pei, J. and N. V. Grishin. 2007. PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* 23: 802–808.

Pei, J., B.-H. Kim, and N. V. Grishin. 2008. PROMALS3D: A tool for multiple sequence and structure alignment. *Nucleic Acids Research* 36: 2295–2300.

Pereira do Lago, A., I. Muchnik, and C. Kulikowski. 2005. A sparse dynamic programming algorithm for alignment with non-overlapping inversions. *Theoretical Informatics and Applications* 39: 175–189.

Petersen, G., O. Seberg, L. Aagesen, and S. Frederiksen. 2004. An empirical test of the treatment of indels during optimization alignment based on the phylogeny of the genus *Secale* (Poaceae). *Molecular Phylogenetics and Evolution* 30: 733–742.

Phillips, A. J. 2006. Homology assessment and molecular sequence alignment. *Journal of Biomedical Informatics* 39: 18–33.

Phillips, A., D. Janies, and W. Wheeler. 2000. Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution* 16: 317–330.

Phuong, T. M., C. B. Do, R. C. Edgar, and S. Batzoglou. 2006. Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic Acids Research* 34: 5932–5942.

Pruesse, E., J. Peplies, and F. O. Glöckner. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28: 1823–1829.

Redelings, B. D. and M. A. Suchard. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evolutionary Biology* 7: 40.

Redelings, B. D. and M. A. Suchard. 2009. Robust inferences from ambiguous alignments. Pp. 209–270 in *Sequence alignment: Methods, models, concepts, and strategies*, ed. M. S. Rosenberg. Berkeley: University of California Press.

Remane, A. 1952. *Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik*. Leipzig: Geest und Portig.

Rieppel, O. 2007. The nature of parsimony and instrumentalism in systematics. *Journal of Zoological Systematics and Evolutionary Research* 45: 177–183.

Roy, A., B. Taddese, S. Vohra, P. K. Thimmaraju, C. J. R. Illingworth, L. M. Simpson, K. Mukherjee, C. A. Reynolds, and S. V. Chintapalli. 2014. Identifying subset errors in multiple sequence alignments. *Journal of Biomolecular Structure & Dynamics* 32: 364–371.

Sahraeian, S. M. E. and B.-J. Yoo. 2011. PicXAA-R: Efficient structural alignment of multiple RNA sequences using a greedy approach. *BMC Bioinformatics* 12: S38.

Sammeth, M. and J. Stoye. 2006. Comparing tandem repeats with duplications and excisions of variable degree. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 3: 395–407.

Sankoff, D., C. Morel, and R. J. Cedergren. 1973. Evolution of 5S RNA and the non-randomness of base replacement. *Nature* 245: 232–234.

Schöniger, M. and M. S. Waterman. 1992. A local algorithm for DNA sequence alignment with inversions. *Bulletin of Mathematical Biology* 54: 521–536.

Seibel, P. N., T. Müller, T. Dandekar, J. Schultz, and M. Wolf. 2006. 4SALE – a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics* 7: 498.

Shapiro, B. A., Y. G. Yingling, W. Kasprzak, and E. Bindewald. 2007. Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology* 7: 157–165.

Shatsky, M., R. Nussinov, and H. J. Wolfson. 2006. Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins: Structure, Function, and Bioinformatics* 62: 209–217.

Sievers, F., D. Dineen, A. Wilm, and D. G. Higgins. 2013. Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics* 29: 989–995.

Simmons, M. P. 2004. Independence of alignment and tree search. *Molecular Phylogenetics and Evolution* 31: 874–879.

Simmons, M. P. and J. V. Freudenstein. 2003. The effects of increasing genetic distance on alignment of, and tree construction from, rDNA internal transcribed spacer sequences. *Molecular Phylogenetics and Evolution* 26: 444–451.

Simmons, M. P., K. F. Müller, and A. P. Norton. 2010. Alignment of, and phylogenetic inference from, random sequences: The susceptibility of alternative alignment methods to creating artifactual resolution and support. *Molecular Phylogenetics and Evolution* 57: 1004–1016.

Simmons, M. P., K. F. Müller, and C. T. Webb. 2008. The relative sensitivity of different alignment methods and character codings in sensitivity analysis. *Cladistics* 24: 1039–1050.

Simmons, M. P., H. Ochoterena, and J. V. Freudenstein. 2002. Amino acid vs. nucleotide characters: Challenging preconceived notions. *Molecular Phylogenetics and Evolution* 24: 78–90.

Simossis, V. A. and J. Heringa. 2005. PRALINE: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research* 33: W289–W294.

Sreeskandarajan, S., M. M. Flowers, J. E. Karro, and C. Liang. 2014. A MATLAB-based tool for accurate detection of perfect overlapping and nested inverted repeats in DNA sequences. *Bioinformatics* 30: 887–888.

States, D. J. and M. S. Boguski. 1991. Homology and similarity. Pp 89–157 in *Sequence analysis primer*, eds. M. Gribskov and J. Devereux. New York: Oxford University Press.

Subramanian, A. R., M. Kaufmann, and B. Morgenstern. 2008. DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biology; AMB* 3: 6.

Tabei, Y., H. Kiryu, T. Kin, and K. Asai. 2008. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics* 9: 33.

Taneda, A. 2010. Multi-objective pairwise RNA sequence alignment. *Bioinformatics* 26: 2383–2390.

Terekhanova, N. V., G. A. Bazykin, A. Neverov, A. S. Kondrashov, and V. B. Seplyarskiy. 2013. Prevalence of multinucleotide replacements in evolution of primates and *Drosophila*. *Molecular Biology and Evolution* 30: 1315–1325.

Thompson, J. D. and O. Poch. 2005. Sequence alignment. In *Encyclopedia of life sciences*. New York: John Wiley and Sons.

van Noorden, R., B. Maher, and R. Nuzzo. 2014. The top 100 papers: *Nature* explores the most-cited research of all time. *Nature* 514: 550–553.

Vellozo, A. F., C. E. R. Alves, and A. Pereira do Lago. 2006. Alignment with non-overlapping inversions in O(n^3)-time. *Lecture Notes in Computer Science* 4175: 186–196.

Vingron, M. and A. von Haeseler. 1997. Towards integration of multiple alignment and phylogenetic tree construction. *Journal of Computational Biology* 4: 23–34.

Wagner, G. P., ed. 2001. *The character concept in evolutionary biology*. San Diego: Academic Press.

Wallace, I. M., G. Blackshields, and D. G. Higgins. 2005. Multiple sequence alignments. *Current Opinion in Structural Biology* 15: 261–266.

Wallace, I. M., O. O'Sullivan, D. G. Higgins, and C. Notredame. 2006. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research* 34: 1692–1699.

Wang, N., E. L. Braun, and R. T. Kimball. 2012. Testing hypotheses about the sister group of the Passeriformes using an independent 30 locus dataset. *Molecular Biology and Evolution* 29: 737–750.

Warburton, P. E., J. Giordano, F. Cheung, Y. Gelfand, and G. Benson. 2004. Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Research* 14: 1861–1869.

Warnow, T. 2013. Large-scale multiple sequence alignment and phylogeny estimation. Pp 85–146 in *Models and algorithms for genome evolution*, eds. Chauve C., N. El-Mabrouk and E. Tannier. London: Springer.

Westesson, O., L. Barquist, and I. Holmes. 2012a. HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics* 28: 1170–1171.

Westesson, O., G. Lunter, B. Paten, and I. Holmes. 2012b. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS ONE* 7: e34572.

Wheeler, W. C. and D. S. Gladstein. 1994. MALIGN: A multiple sequence alignment program. *The Journal of Heredity* 85: 417–418.

Wheeler, W. C., N. Lucaroni, L. Hong, L. M. Crowley, and A. Varón. 2015. POY version 5: Phylogenetic analysis using dynamic homologies under multiple optimality criteria. *Cladistics* (in press).

Wilm, A., D. G. Higgins, and C. Notredame. 2008. R-Coffee: A method for multiple alignment of non-coding RNA. *Nucleic Acids Research* 36: e52.

Wray, G. A. and E. Abouheif. 1998. When is homology not homology? *Current Opinion in Genetics & Development* 8: 675–680.

Yoshizawa, K. 2010. Direct optimization overly optimizes data. *Systematic Entomology* 35: 199–206.

Zhang, Y. 2008. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology* 18: 342–348.

Zola, J., X. Yang, S. Rospondek, and S. Aluru. 2007. Parallel T-Coffee: A parallel multiple sequence aligner. Pp. 248–253 in *Proceedings of the ISCA 20th international conference on parallel and distributed computing systems (PDCS-2007)*, eds. G. Chaudhry and S.-Y. Lee. Cary: International Society for Computers and Their Applications.