Plant Syst Evol (2009) 282:127–149
DOI 10.1007/s00606-008-0072-5

REVIEW

A framework for phylogenetic sequence alignment

David A. Morrison

Received: 23 April 2007 / Accepted: 1 November 2007 / Published online: 30 July 2008

Abstract A phylogenetic alignment differs from other forms of multiple sequence alignment because it must align homologous features. Therefore, the goal of the alignment procedure should be to identify the events associated with the homologies, so that the aligned sequences accurately reflect those events. That is, an alignment is a set of hypotheses about historical events rather than about residues, and any alignment algorithm must be designed to identify and align such events. Some events (e.g., substitution) involve single residues, and our current algorithms can successfully align those events when sequence similarity is great enough. However, the other common events (such as duplication, translocation, deletion, insertion and inversion) can create complex sequence patterns that defeat such algorithms. There is therefore currently no computerized algorithm that can successfully align molecular sequences for phylogenetic analysis, except under restricted circumstances. Manual re-alignment of a preliminary alignment is thus the only feasible contemporary methodology, although it should be possible to automate such a procedure.

Keywords Molecular sequences · Sequence alignment · Phylogenetic analysis

It is not difficult to find publications where the authors have used three different tree-building methods (e.g., with parsimony, likelihood and posterior probability as the respective optimality criteria), but it is much more difficult to find publications where more than one alignment method has been used (e.g., Prychitko and Moore 2003), in spite of the fact that it has been repeatedly shown that alignments have at least as much effect as tree-building on the outcome of the phylogenetic analysis (Ellis and Morrison 1995; Morrison and Ellis 1997; Beebe et al. 2000; Mugridge et al. 2000; Quandt et al. 2003; Hertwig et al. 2004; Gillespie et al. 2005; Ogden and Rosenberg 2006; Martin et al. 2007). I conclude from this that many researchers believe that an alignment can be taken as “fixed”, and that our current methods are capable of producing useful fixed alignments. I contend that this attitude is seriously mistaken, except under specific circumstances.

Instead, I argue that our current procedures for the alignment of multiple molecular sequences are mis-directed at quite a fundamental level. Indeed, it can be argued that no-one has yet presented reasonable theoretical principles for phylogenetic sequence alignment. Our current procedures have as their goal the alignment of residues, such as nucleotides or amino acids, so that the ensuing alignment is seen to be a set of hypotheses about the residues. Here, I make the case that in order to be useful for phylogenetic analysis an alignment must be a set of hypotheses about the events that led to the sequence patterns, rather than about the patterns themselves. That is, a phylogenetic alignment should be seen as aligning evolutionary events rather than as aligning molecular residues. This might be called an event-based sequence alignment.

An alignment is a data matrix, and a phylogenetic tree is simply a re-representation of that data matrix (Mishler 2005). Therefore, every alignment, as well as every tree, should explicitly reflect evolutionary history if it is to be part of a phylogenetic analysis. The key to a successful phylogenetic analysis is the care with which the data matrix

D. A. Morrison (✉)
Department of Parasitology (SWEPAR), National Veterinary
Institute and Swedish University of Agricultural Sciences,
751 89 Uppsala, Sweden
e-mail: David.Morrison@bvf.slu.se

has been evaluated for potential homology (Mishler 2005), which for an alignment means evaluating the scenario of events that is being proposed as having created the pattern at each aligned position. If the alignment unambiguously represents those events then the subsequent tree-building will be straightforward.

Sometimes, aligning the residues will align the events (e.g., when the events are substitutions) but often they will not (e.g., when the events are duplications or inversions). The focus should be on the events, and thus the sequence blocks involved in those events, not on the similarity of individual nucleotides (or amino acids). There is no known algorithm for aligning the products of unobservable historical events, and so none of our current alignment procedures can be assured of producing an alignment that is useful for phylogenetic purposes.

The particular issues that I discuss here are not unique to non-coding sequences, but they do come into much sharper focus when considering suitable procedures for the alignment of such sequences. The creation of a biologically relevant alignment of protein-coding sequences, for example, is much more amenable to our current strategies than is that of non-coding sequences (Creer 2007). For the purposes of this paper I have found it useful to distinguish three types of sequence: sequences that code for proteins (protein-coding sequences), sequences that code for structural/functional RNAs such as rRNAs and tRNAs that are involved in protein expression (RNA-coding sequences), and conserved non-genic sequences (non-coding sequences).

There have been a number of published papers where the authors have considered individual phylogenetic alignments, without necessarily discussing general principles. These include the works of Cammarano et al. (1999) and Lebrun et al. (2006) for protein-coding sequences, and Kjer et al. (1994), Kjer (1995) and Gillespie (2004) for RNA-coding sequences. For non-coding sequences, general principles have been listed by a number of authors, including Golenberg et al. (1993), Kelchner and Clark (1997), Hoot and Douglas (1998), Graham et al. (2000), Borsch et al. (2003) and Löhne and Borsch (2005). Here, I try to provide a general theoretical framework for molecular sequence alignment that integrates all of these ideas.

Alignment and homology

Two sequences are homologous if they have descended through a chain of replication from a common precursor molecule (Cartmill 1994), and residues are homologous if they have maintained the same positions in those sequences (Dewey and Pachter 2006). We are often told that alignment of molecular sequences should, in some way, be related to hypotheses of homology regarding the evolutionary origin of those sequences. However, this is not necessarily true except when the sequence alignment is to be used for a phylogenetic analysis. At heart, an alignment is simply a preparatory way of arranging the data for analysis, and the best arrangement depends on the purpose of the analysis. There are actually at least four distinct purposes for constructing a multiple sequence alignment that can be considered to be biologically relevant (Morrison 2006), and for only one of these is homology essential.

The four different purposes for sequence alignment are: (1) database searching, (2) structure prediction, (3) sequence comparison, and (4) phylogenetic analysis. For these, the analysis objectives are, respectively: (1) to maximize the distinction between homologous and non-homologous sequences, (2) to deduce the secondary and tertiary structure of a gene product from knowledge of the gene sequence, (3) to juxtapose residues representing conserved sequence features (e.g., conserved motifs, such as occur at active sites), and (4) to produce plausible hypotheses of evolutionary homology among the sequence residues. These distinct purposes are not exactly uncommon in molecular biology. For example, two of the most-cited publications in the biological sciences describe sequence-alignment computer programs: BLAST for pairwise database searching (>40,000 ISI Web of Science citations as of April 2007) and Clustal for multiple sequence comparison (>33,000 citations).

Most of the alignment computer programs were originally developed for sequence comparison, and they were later applied to the other three purposes in an ad hoc manner, without regard for their suitability. Fortunately, specialist programs have recently been developed for structure alignment, database-search alignment, and the search for functionally conserved subsequences (Morrison 2006). Unfortunately, little attention has been paid to the development of computer programs specifically for multiple alignment in the context of phylogenetic analysis.

The main practical difference between these various types of alignment is that homology of the aligned residues is optional for all of them except for phylogenetic analysis. Homology can be helpful for structure prediction, database searching and (especially) sequence comparison, but their respective objectives can often be achieved without it. For example, a shift in function from one residue of a sequence to its immediate neighbor may mean that the optimal alignment for structure prediction aligns the functionally equivalent residues rather than the historically equivalent ones. Phylogenetic analysis, on the other hand, requires that the historically equivalent residues must always be aligned.

Homology involves defining characters and their states. For phenotypic characters this is often straightforward, although it can be confusing in practice. For example, bracts, bracteoles, sepals, petals, nectaries, anthers and ovaries are all modified leaves (i.e., during their evolutionary history they have been modified from leaves into their current form); and it can be complex defining the characters and their alternative states just by looking at contemporary organisms, because there are a very large number of types of units (modified leaves) to compare and arrange (into the characters bracts, bracteoles, etc.). The situation is both better and worse for DNA sequences. It is better in the sense that the units being compared (the nucleotides A, C, G, T) are few and easy to recognize; but it is worse in the sense that it is difficult to arrange the units into characters (the "columns" in the standard way of arranging an alignment) and their states (which nucleotides go in which columns), precisely because the units all look the same. That is, for phenotypic characters there are lots of units to compare, and the units themselves provide clues as to which character they are part of, but for genetic data the units are identical, and thus provide no intrinsic clues. The units must therefore be shuffled back and forth among the characters, trying to work out which combination of characters and states represents the evolutionary history of the units. In this sense, the distinction between characters and character states is rather vague, as characters are simply hypotheses of homology at a more inclusive level than those of character states (Patterson 1988).

The idea of homology is not different in any fundamental way for different sequence types, whether they are protein-, RNA- or non-coding. Here, the expression "non-coding" covers a multitude of sequence types, such as inter-genic regions, transcribed spacers, introns, the many types of microRNAs and snoRNAs, transposable elements, the mitochondrial control region, cis-regulatory sequences, and other sequences involved in regulating gene expression, such as promoters and enhancers. However, as far as the practical business of sequence alignment is concerned, the important distinction is between conserved sequences and non-conserved sequences. Variation in sequence conservation results from variation in functional and structural constraints, and reduced conservation usually leads to length variation as a result of microstructural changes. Length variation, in turn, leads to multiple equally optimal alignments, although increased rates of substitution are also observed. Protein-coding and RNA-coding sequences are, in general, more highly conserved than are non-coding sequences, which is thus the only notable distinction between them as far as aligning the sequences is concerned. However, conserved coding regions have been reported to constitute only 1–20% of the genome of multicellular eukaryotes (Szymanski et al. 2007), and highly conserved non-coding sequences about twice this, and so there is considerable scope for alignment of non-coding sequences in the rest of the genome.

Proposing and testing homologies

Homology assessment can be considered to involve two steps (de Pinna 1991). The first step is the conjecture, prior to data analysis, that similarity among certain characters and character states may represent evidence of evolutionary groupings of the taxa; this is primary homology. The second step concerns the recognition of congruence among the primary homologies as a result of a tree-building analysis of the data—the shared derived character states (synapomorphies) on the phylogenetic tree represent homologies; this is secondary homology. Thus, primary homology is a conjectural assessment of homology prior to phylogenetic analysis (an assessment of essential sameness) while secondary homology is a corroborated homology assessment subsequent to the analysis (an assessment of congruence that explains the sameness). From this perspective, sequence alignment is primary homology assessment (Brower and Schawaroch 1996).

It has been traditional in phylogenetic analyses (i.e., when dealing with phenotypic characters) to keep assessment of primary and secondary homology separate, one being a priori and the other a posteriori with respect to the tree-building procedure. Therefore, it is hardly surprising that alignment and tree-building have been treated as separate activities in molecular biology (Patterson 1988). In particular, testing of homologies is not the only possible goal of a sequence alignment—sequence comparison may be best done in an evolutionary context, for example, (Dobzhansky 1973). Conversely, it is also possible to construct phylogenetic trees without first aligning the sequences, although this is usually less successful (Höhl and Ragan 2007).

However, a strong argument has been presented to treat alignment and tree-building as two sides of the one coin. That is, we should be optimizing the alignment and the tree simultaneously, since they are inter-dependent (Sankoff et al. 1973). This is because an alignment has a built-in phylogenetic structure, and a phylogenetic tree implies a particular alignment, so that the duality obviates the need to estimate them separately. Furthermore, in practice there have often been contradictory assumptions applied to sequence alignment and tree-building in the same phylogenetic analysis. So, in the name of methodological consistency it has been argued that no distinction should be made between assessment of primary and secondary homology.

This has resulted in the development of two different strands to the same philosophy of sequence alignment, known as direct optimization (Phillips et al. 2000) and statistical alignment (Lunter et al. 2005). The first method directly optimizes ancestral sequences while treating gaps as a fifth character state rather than as missing data. The correct alignment is seen to be the one that produces the minimum-cost phylogenetic tree, where all of the cost parameters (substitution costs, gap penalties, sequence weights, etc.) are specified concurrently for both the alignment and the tree. Here, the idea of "cost" has been implemented in the POY computer program in the form of both parsimony analysis (Wheeler 1996) and likelihood analysis (Wheeler 2006).

Statistical alignment, on the other hand, 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, and some criterion is then used to optimize the parameters in relation to the model, such as either maximizing the likelihood or the Bayesian posterior probability. To date, two versions for multiple sequences have been implemented, in the AliFritz (Fleissner et al. 2005) and BAli-Phy (Redelings and Suchard 2005) computer programs.

As I will show in a later section using an example, direct optimization and statistical alignment can lead to quite different alignments from the alternative methods, particularly for non-coding DNA. They make defining characters and their states much more complex, because the final tree plays a part in defining the characters and their states. All alignment methods shuffle character states among characters as they proceed, with the implicit objective of defining the characters (the residue columns in the alignment). Shuffling the character states while simultaneously building the phylogenetic tree means that congruence among the characters becomes part of the definition of the characters rather than a test of them. A specific example is shown in Fig. 1.

If nothing else, this can create artifacts as a result of the inter-play of alignment and tree-building. For example, a set of ambiguously aligned characters (i.e., where there are several equally optimal alternative alignments) can be made congruent with a single unambiguously aligned character, resulting in an apparently well-supported unambiguous alignment (Simmons 2004). Furthermore, neither direct optimization nor statistical alignment, in their current implementations, has any means to detect whether all of the sequence regions being aligned have the same evolutionary history (i.e., they insist upon a single tree for the entire alignment). Both methods can deal with "indivisible" sequence blocks but neither has any effective way to define those blocks, because different histories are not built into their models. Thus, direct optimization and statistical alignment can sacrifice biological plausibility in their attempts at methodological consistency (i.e., applying the same method to both alignment and tree-building).

Moreover, an hypothesis and its test must be kept independent of each other, otherwise there is no "test". Direct optimization and statistical alignment make the optimization problem the purpose of the exercise (thus confounding descriptive and ontological parsimony; Simmons 2004), rather than the purpose being the proposing and testing of phylogenetic (homology) hypotheses. To introduce an analogy, this is like giving a group of students a set of exam questions, and then adjusting each question for each individual answer, so that the students all score 100% (perhaps leading to the conclusion that the teacher is very able and the students are very intelligent). Alignments and trees are linked (as are questions and answers), but that does not mean we must make them totally inter-dependent. The two procedures can be kept separate so that the results can be treated as tests.

For multiple genes, each gene represents a potential test of both homology (alignment) and phylogeny (tree). Congruence among the genes can be considered to be strong evidence for both the alignment and the tree. If we simultaneously optimize all of the genes then we lose both

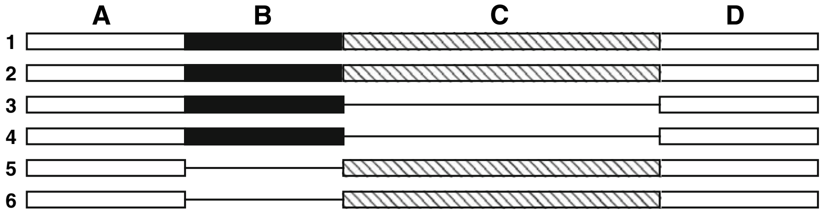

Fig. 1 An artificial alignment illustrating some of the potential problems both with progressive alignment and with simultaneous alignment and tree-building. Regions A–D are shared in various combinations between sequences 1–6; single lines represent sequence absence. This pattern might be created, for example, by multiple isoforms of alternatively spliced gene products. A progressive alignment would first align sequences 1 & 2, then 5 & 6, and then 3 & 4, all pairs without gaps. Then it would align 3 + 4 with 5 + 6, inserting gaps into region B of sequences 3 + 4 to align it with region C of 5 + 6. Finally, it would align 1 + 2 with 3 + 4 + 5 + 6, aligning region C. Thus, sequences 3 and 4 will be mis-aligned with respect to sequences 1 and 2. Golubchik et al. (2007) show that this is precisely what ClustalW does, for example. A simultaneous alignment and tree-building analysis would recognize that the tree specified by region B, which unites sequences 1–4, differs from the tree specified by region C, which unites sequences 1, 2, 5 & 6. Because region C has more data than region B, the tree specified by region C is better supported, and so the alignment of region B will be adjusted to match the tree specified by region C. Region B may thus be mis-aligned

of these tests. This is because the tree topology supported by one gene tree can influence the alignment of another gene (Simmons 2004). A second data set is then not being used to independently test the tree supported by the first data set, but is instead merely being assessed for its degree of congruence with that tree.

Thus, an alignment from direct optimization or statistical alignment is not a primary hypothesis to be tested but is, instead, a hypothesis that has already been tested on a tree (a confirmed hypothesis). If we wish to see the primary hypotheses of homology in order to evaluate their biological plausibility (e.g., different sequence regions might have different histories), then we need to see an alignment. The framework that I am presenting here assumes that alignment and tree-building are separate issues, and that we intend to develop an alignment that is independent of its subsequent testing on a tree. Thus, proposing a hypothesis is distinct from testing it.

In practice, hypotheses can be generated in any manner at all, but clearly we are interested in generating "useful" ones in a phylogenetic context. We therefore need independent sources of evidence for potential homologies. Comparative analysis has been the traditional way to acquire this evidence, and it is straightforward to apply this approach to sequences as well (Morrison 2006), based on the underlying molecular processes that lead to the changes associated with the homologies. Thus, defining "alignment events" is similar to defining morphological character states, and to defining transformation series between those states.

Homologies and events

Homologies arise as the result of one or more events in evolutionary history. That is, some event occurs that changes an ancestral character state into a derived character state, and it is the sharing of the derived character state that represents the homology. From this point of view, it is the ability to conceive of the event that allows us to recognize the potential homology.

The theory of multiple sequence alignment for phylogenetics is thus to identify the events that have occurred in history, while the practice is to align the sequences so that the history of the events is evident. This practice involves first searching for evidence of the events and the bits of sequence involved, and then representing the individual events in the best way (e.g., making sure that separate events are not aligned against each other, being consistent about the representation, etc.).

This is exactly the opposite approach to multiple alignment to what has traditionally been done. Here, the events are identified as the alignment procedure proceeds, whereas traditionally one identifies the events only *after* the alignment has been produced (Kim and Sinha 2007). That is, the events have been treated as being a conclusion from the alignment rather than being a cause of it.

These events involve known molecular mechanisms, such as slippage during DNA replication/repair, small inversions and deletion of loop regions in DNA secondary structure (for small sequence blocks), as well as chromosomal processes such as recombination, gene conversion and horizontal gene transfer (for large sequence blocks). From the practical point of view, it is worthwhile recognizing two types of event: (1) those that can be detected within a single sequence; and (2) those that can be identified only by comparing two (or more) sequences. The most common events of type (1) are duplications (copying of a subsequence to another location), notably tandem repeats (copying to an immediately adjacent position; see Fig. 2) and inverted repeats (reverse-complementing the copy; see Fig. 2), because they involve copies of a region within the same sequence. The most common events of type (2) are substitutions (replacement of one nucleotide by another), inversions (replacement of a subsequence by its reverse complement; see Fig. 3), translocations (removal of

```
              Source of repeat    Inverted repeat          Repeat              Source of inverted
                                                                                     repeat
Wa-S   ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Fl-1S  ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Af-S   ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Fr-S   ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Fl-2S  ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Ja-S   ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Fl-F   ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Ja-F   ATAAA|AATAAACCATAAACTAGGC|------------------|-------------------|AGCGCT|GCCGTCGCCGTCTGAGCA|GCCTG
Fr-F   ATAAA|AATAAACCATAAACTAGGC|TGCTCAGCCGGCGACGGC|AATAAACCATAAACTAGGC|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Wa-F   ATAAA|AATAAACCATAAACTAGGC|TGCTCAGCCGGCGACGGC|AATAAACCATAAACTAGGC|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
Af-F   ATAAA|AATAAACCATAAACTAGGC|TGCTCAGCCGGCGACGGC|AATAAACCATAAACTAGGC|AGCGCT|GCCGTCGCCGGCTGAGCA|GCCTG
```

Fig. 2 A gapped section in the sequence alignment of the inter-genic region preceding the *Adh* gene of 11 strains of *Drosophila melanogaster*, from Kreitman (1983). This shows that two distinct events, a repeat and an inverted repeat, have created the apparent single insertion. The vertical bars delimit the various annotated regions, while the underlined nucleotides indicate those parts of the sequence that are capable of pairing to form a secondary-structure stem. Note that it is also possible, for the first eight sequences, to move the block of subsequences from the "Source of repeat" region to the "Repeat" region, since there is no obvious evidence here to distinguish which is the template and which the copy—they have been left-aligned as a convention

```
                               Source of      Inversion      Inverted
                                 repeat                        repeat
Utricularia rigida      TCGAAT|CGTTCCAA|AACCTTGTTTGAATTC|TTGGAACA|TCTAAA
Utricularia nana        TCGGAT|TGTTCCAA|AACCTTGTTCGCATTC|TTGGAGCG|TAAAAT
Utricularia spiralis    TCGGAT|CGTTCCAA|GGATTCGAACAAGGTT|TTGGAACG|TAAAAA
Utricularia uliginosa   TCGGAT|CGTTCCAA|GGATTCAAACAAGGTT|TTGGAACG|TAAAAA
Utricularia foveolata   TCGGAT|CGTTCCAA|GAATTCAAACAAGGTT|TTGGAACC|TAAAAA
```

Fig. 3 An ungapped region in the sequence alignment of the *trn*K intron of five species of *Utricularia* (Lentibulariaceae), from Müller and Borsch (2005a). This shows two possible events, an inverted repeat and an inversion, with the last three sequences being inverted (i.e., reverse complemented) with respect to the first two sequences. The *vertical bars* delimit the various annotated regions, while the *underlined nucleotides* indicate those parts of the sequence that are capable of pairing to form a secondary-structure stem. Here, the inversion plus two substitutions (i.e., three events) has created an apparent set of 12 substitutions (i.e., 12 events) in the 16 aligned nucleotides of the inverted region. Müller and Borsch (2005a) reverse-complemented the inverted region before phylogenetic analysis, but it could also have been treated as a staggered alignment (as was done under similar circumstances by Graham et al. (2000)

a subsequence and its insertion at another location) and transpositions (interchange of subsequences), deletions (complete removal of a subsequence), and insertions (addition of a novel subsequence), all of which leave no traces within a single sequence. It is perhaps worth noting that a sequencing error is also an "event" that leaves evidence behind, although it is a lot more recent event than the biologist is presumably expecting.

Having a list of the events inferred by the alignment, particularly those resulting in length variation, as shown in Fig. 4, explicitly makes the point that the alignment is a series of hypotheses. That is, the alignment is not just a set of hypotheses about homologous residues but is *also* a set of hypotheses about the events that created the differences between the sequences. This information can be considered to be a very valuable part of every alignment (examples are shown by Löhne and Borsch 2005; Müller and Borsch 2005b; Borsch et al. 2007), and it would be handy to have an editor that can annotate the alignment with the postulated events. It might also be useful to have some nomenclatural scheme for the observed patterns created by the inferred events (Kelchner 2000), for easy reference. These suggestions all have clear analogies when defining characters for phenotypic data, where it is traditional to have a list of characters and their states in each phylogenetic analysis along with detailed argumentation for how/why each character was defined a priori.

Note that the term "indel" has little relevance to the discussion here, in spite of its apparent importance in previous discussions of sequence alignment. The term "indel" is usually contrasted with the term "substitution", the former being associated with sequence differences involving length variation and the latter not. However, the term "indel" implies a single concept, and my argument here is that there can be many concepts involved in length variation (e.g., duplication, deletion, insertion). Furthermore, there are several events that do not create length variation (e.g., substitution, inversion) or may not do so (e.g., translocation, transposition). The focus on events rather than residues makes the contrast between "indel" and "substitution" too simplistic.

```
AF288660_MT3663 ATTGCCCAATGTCAAAAAAAAAAGATGAGGCAG
Structure       -<-<<<---<<<<<-------->>>>>->>>>-
AF359481_T1     ATTGCCCAATGTCAAAAAAAA AGATGAGGCAG
AF359482_T2     ATTGCCCAATGTCAAAAAAAA4AGATGAGGCAG
AF359483_T3     ATTGCCCAATGTCAAAAAAAA AGATGAGGCAG
AF359484_T4     ATTGCCCAATGTCAAAAAAAA AGATGAGGCAG
AF359485_T5     ATTGCCCAATGTCAAAAAAAAAA ATGAGGCAG
AF359486_T6     ATTGCCCAATGTCAAAAAAAAAA5ATGAGGCAG
AF359487_T7     ATTGCCCAATGTCAACAAAAAAA ATGAGGCAG
AF359488_T8     ATTGCCCAATGTCAACAAAAAAA ATGAGGCAG
AF359489_T9     ATTGCCCAATGTCAAAAAAAAGAGATGAGGCAG
AF359490_T10    ATTGCCCAATGTCAAAAAAAAGAGATGAGGCAG
AF359491_T11    ATTGCCCAATGTCAAAAAAAAGAGATGAGGCAG
AF359492_T12    ATTGCCCAATGTCAAAAAAAAGAGATGAGGCAG
AF359493_T13    ATTGCCCAATGTCAAAAAAA2GAGATGAGGCAG
AF359494_T14    ATTGCCCA1TGTCAAAAAAA GAGATGAGGCAG
AF359495_T15    ATTGCCCAATGTCAAAAAAA3GAGATGAGGCAG
AF359496_T16    ATTGCCCAATGTCAAAAAAAAGAGATGAGGCAG
```

Fig. 4 Aligned DNA of a single stem from the *Trypanosoma cruzi* SSU rRNA, for 16 sequences from the data set of Sanson et al. (2002). The top row shows the ancestral sequence of the experimentally produced phylogeny. The second line shows the second-order structure of the stem, based on Cannone et al. (2002), with opposite *angle-brackets* indicating paired nucleotides; note the unusual A-A pair at the top of the stem. The authors sequenced all of the ancestors in their experiment, and so the events leading to the alignment can be reconstructed unambiguously. All of the six events are *boxed*, with the deletions numbered. Note that the known history indicates that deletions 2 and 3 are independent. The positions of deletions 1, 2, 3 and 4 are ambiguous, because they occur in homonucleotide blocks. The sequences have been left-aligned for these deletions, so that the gaps are placed as far to the right as possible. This is merely a convention, so that the alignment procedure is objective and repeatable. The position of deletion 4 is partly constrained by the second-order structure (if it is any further to the right then it would disrupt the stem pairing); thus, all of the deletions occur in unpaired positions except for 5. If deletion 5 is moved then this would imply two events (a deletion and a substitution), rather than one, which is a less parsimonious reconstruction

We thus need to adopt a different attitude to multiple sequence alignment: do not align anything unless there is a clear reason to do so. Only residues that we are proposing to be homologous should be aligned, while residues that have no homologues should not be aligned against any other residues (Morrison 2006). For example, if two sequences have insertions relative to the other sequences, then do not align the two insertions unless there is evidence to do so, but instead leave them as staggered alignments (Barta 1997), as shown in Fig. 5. This is the opposite behavior to that of all of the computer programs based on

```
Iphiseius degenerans            CATTTGTTTCAGTATATAAA---------CCGAATC-ATACGTATTTACCTTTGC
Typhlodromus pyri               CATTTGTTTCAGTATATAAA---------CCGTATC-ATAAGTATTTACCTTTGC
Euseius finlandicus             CATTTGTTTCAGTATATAAA---------CCGAATC-ATAAGTATTTACCTTTGC
Euseius concordis               CATTTGTTTCAGTATATAAA---------CCGAATC-ATAAGTATTTACCTTTGC
Neoseiulus cucumeris            CATTTGTTTCAGTATATAAA---------CCGTATC-ATACGTATTTACCTTTGC
Neoseiulus fallacis             CATTTGTTTCAGTATATAAA---------CCGTATC-ATACGTATTTACCTTTGC
Neoseiulus californicus         CATTTGTTTCAGTATATAAA---------CCGTATC-ATACGTATTTACCTTTGC
Metaseiulus occidentalis        CATTTGTTTCAGTATATAAT---------CCGTATC-ATACGTATTTACTGTTGC
Tinaminyssus sartbaevi          CACTTGTTTCAGTATATTAAA---------CGAAGCAATAAGTAATTACTATTGC
Tinaminyssus streptopelioides   CATTTGTTTCAGTATAAAAAA----------GTACTAATGCGTAGTTACTATTGC
Tinaminyssus melloi             CATTTGTTTCAGTATACAAAA---------C-TAGCATAAAGTAGTTACTATTGC
Tinaminyssus bubulci            CATTTGTTTCAGTATACAAA----------C-AAGCAATACGTAATTACTATTGC
Tinaminyssus minisetosum        CACTTGTTTCAGTATATTGA----------CCTAGCAATACGTAGTTACTATTGC
Tinaminyssus columbae           CACTTGTTTCAGTATATTGA----------CATAGAAATACGTAGTTACTATTGC
Sternostoma strandtmanni        CATTTGTTTCAGTATATAAA-AGA------TGT-CCAATACGTAATTGCTGTTGC
Sternostoma boydi               CATTTGTTTCAGTATATAAA-AGA------TGT-CCAATACGTAATTGCTGTTGC
Sternostoma fulicae             CATTTGTTTCAGTATATATA-AAA------TGT-TCAATACGTACTTGCTGTTGC
Sternostoma turdi               CATTTGATTCAGTATATAGA----C-----TGT-AGAATACGTAGTTACTATTGC
Tropilaelaps koenigerum         CACTTGTTTCAGTATATAA------CT---CGT-CGTATAAGTACTGACTATTGC
Tropilaelaps clareae            CACTTGTTTCAGTATATAA------CT---CGT-AGTATATGTACTTACTATTGC
Rhinonyssus tringae             CAGTTGTTTCAGCATATGA--------GG-AGT-ACATTACGTAGTTACTATTGC
```

Fig. 5 Aligned DNA sequences at the boundary of the ribosomal 5.8S–ITS2 region of 21 species of mites (Acari; Mesostigmata), from the data set of Morrison (2006). The non-stem region is length-variable, but only putatively homologous nucleotides have been aligned. This is an example of a staggered alignment, in which alignment of a pair of residues is an explicit statement that they are hypothesized to be homologous, and non-homologous residues not aligned (i.e., they are staggered with respect to each other). A maximum-similarity alignment would align many of the blocks in the gapped region that are not aligned here, for example aligning many of the As and Cs into single columns. The species are arranged in approximate taxonomic order, illustrating that the various insertions here are usually phylogenetically informative

sequence similarity, which only unalign residues when similarity falls below a certain level (determined by the gap costs). For a phylogenetic alignment, we should not align sequences unless there is clear evidence of homology, just as we do not hypothesize homology of phenotypic structures unless there is good evidence from comparative analysis for doing so. If nothing else, this minimizes the number of false positives (i.e., incorrect synapomorphies in the phylogenetic tree).

Hypotheses about evolutionary events should be plausible and parsimonious. Plausibility is an obvious requirement for any hypothesis, and this is where our biological knowledge plays its primary role in sequence alignment. As our knowledge about the relevant molecular mechanisms changes, so will our assessment of the relative plausibility of our hypotheses under specified circumstances.

Parsimony, in the philosophical sense, is a methodological tool that says "in the absence of evidence to the contrary, choose the simplest explanation". People often overlook the initial caveat in this definition, and make parsimony the goal instead of leaving it as a convention. However, parsimony cannot be the objective unless evolution acts parsimoniously (i.e., ontological parsimony sensu Johnson 1982), which is contrary to much of our empirical evidence. My use of parsimony here is thus strictly a methodological tool, where we prefer simpler explanations until we accumulate evidence that reality is more complex (i.e., descriptive parsimony). My argument here is that we have evidence that the substitution/indel dichotomy is too simple to be useful, and thus we should be looking for a more complex description of alignment; but we must still employ the principle of parsimony in searching for that new description (e.g., Fig. 4).

Exactly how to quantify a parsimony score in practice is a separate issue that has never been satisfactorily resolved, leaving us with various optimality criteria for simplicity, such as parsimony score and likelihood. The strength of so-called "parsimony analysis" is that it is based on the observed data alone ("what you see is all there is"), which is somewhat akin to conditional analyses in statistics. Its weakness is that biases in the data are translated directly into biases in the resulting analysis. For "model-based analyses", such as those based on likelihood, the strength is that biases can be corrected by the model. The weakness is that the models are usually unrealistic, and we do not yet know enough about how realistic they need to be in order to produce useful results. I am not specifically concerned here with how to quantify parsimony, but I will briefly return to the issue below.

The traditional approach to alignment is to put gaps into the sequences and then to infer the events from the resulting multiple alignment. I am proposing to reverse this process, and to first infer the events and then to insert gaps to represent those events. Thus, we need to have alternative evidence for the events, because we do not yet have the multiple alignment, which is the traditional evidence.

Alignment evidence

Put simplistically, the objective is to align a set of unobservable historical events. These events have left traces in the contemporary sequences, such as observable micro-structural changes. Can we use these (observable) traces to reconstruct the (unobservable) events, so that we can construct the phylogenetic tree? Unfortunately, there cannot be an algorithm for aligning unobservable historical events, and therefore we must use whatever evidence is at hand with regard to the mutational mechanism in any particular situation. This might come from the primary structure of the nucleotide sequences themselves, either as patterns in single sequences or as motifs shared among sequences, or it might come from the secondary or tertiary structure of any molecule that the sequences code for.

This approach relies on a model that defines the types of evolutionary events that can be expected. We may need different methods tailored to different data types, such as protein-coding sequences, RNA-coding sequences or non-coding sequences, as these have different functional constraints and sequence characteristics. Even within these sequence types there may be different expectations for different sequence regions. This is because different events occur with different frequencies in different molecular structures (Borsch and Quandt this volume), and our ability to detect such events differs. For example, length variation is less likely to occur when there are paired nucleotides or other rigid structural features, and therefore different alignment strategies may be needed for different structures.

The events that we are investigating include: substitutions, duplications (notably tandem repeats and inverted repeats), inversions, translocations, deletions and insertions. Whenever two sequences differ at a particular location it will be the result of one of these events, and our objective is to work out which combination of these events is involved along the whole length of the sequences. That is, we need to construct a detailed historical scenario that turns a single ancestral sequence into the set of contemporary sequences.

The evidence for these events can include, but is not restricted to: (1) pairwise sequence similarity; (2) sequence sub-groups with high similarity; (3) motifs conserved across the sequences; (4) identification of within-sequence patterns; (5) a previous phylogenetic tree or current taxonomic hypothesis; (6) structure of the encoded products (if any); and (7) failure of database searches to detect close matches. Each of these sources of evidence can be investigated independently, and automated methods exist for many of them.

Sequence similarity (evidence type 1) is the classical approach taken by all current computerized algorithms. Similarity is strong evidence of homology (Patterson 1988), but it is important to note that it is the *pattern* of similarity that provides the evidence for phylogenetic history, not the similarity per se. That is, phylogenetic events create a specific set of similarity patterns, and it is the set as a whole that is important. For example, it is the fact that unique motifs are shared across several sequences that is of importance, whereas the current progressive-alignment programs look at the sequences only pairwise (or three at most; Colbourn and Kumar 2007), as shown in Fig. 1.

It is for this reason that sequence sub-groups with high similarity are an important source of evidence (evidence type 2). These subgroups are classically found using local alignment algorithms, which is why there has been much interest in combining global and local alignment strategies (see Morrison 2006). Having found an alignment of the subgroups, perhaps the other sequences can then be fitted to this partial alignment. Indeed, this strategy is adopted by all of the current genome-alignment algorithms (Pollard et al. 2004). This source of evidence is thus likely to be among the most important for detecting those events that cannot be detected in a single sequence (e.g., inversion, translocation). For example, examination of a pairwise dot plot can easily show regions where there have been translocations, duplications and inversions in one sequence with respect to the other (see Brudno et al. 2003b).

For the same reasons, searching for conserved motifs (evidence type 3) can be productive (Grundy and Noble 1999), such as occur particularly in functional regions. Functional constraints lead to evolutionary constraints, which lead to a relative lack of events to detect when aligning sequences. Sometimes only a motif is conserved at the sequence level even if the secondary structure of the encoded product is highly conserved. Therefore, conserved functional motifs, whether in protein-coding (e.g., disulphide bridges), RNA-coding (e.g., conserved single-stranded regions) or non-coding (e.g., Shine-Delgarno motifs) sequences, can help us to get a preliminary alignment correct, so that the events themselves can then be identified. This approach has recently been adopted for protein-coding sequences (Du and Lin 2007; Papadopoulos and Agarwala 2007), where the computer programs use databases such as CDD and PROSITE as the source of motif information, which is then used to constrain the multiple alignment.

Large and well-conserved regions can be detected easily by local alignment methods (Kumar and Filipski 2007), but poorly conserved motifs are harder to find and require specialist approaches, with different strategies being applied to protein-coding and RNA-coding sequences (e.g., Frith et al. 2004; Yao et al. 2006), since their motifs have different characteristics, which of these strategies is more likely to be appropriate for non-coding sequences will be determined by the type of non-coding sequence at hand; for example, introns and transcribed spacers, which contain conserved sites functioning in correct splicing, usually have similar motifs to those of RNA-coding sequences (Kelchner 2002).

Identification of within-sequence patterns (evidence type 4), such as repeats, seems to be rarely done in any explicit way by phylogeneticists. However, it is an area in which much automated help is available (see Morrison 2006 for a list of available programs), which is convenient because tandem repeats are usually reported to be the most common form of “indel” (e.g., Messer and Arndt 2007), and in my experience they are the phenomena most likely to mislead similarity-based alignment programs. Having located these patterns, you then have to check whether they are variable across the sequences, of course, and O’Dushlaine and Shields (2006) offer some automated help for such comparisons.

One useful heuristic strategy for identifying events (evidence type 5) is to arrange the sequences into their current taxonomic groups or the groups indicated on some previous phylogenetic tree (Prychitko and Moore 2003), because similar alignments can be expected for closely related groups. This approach can often highlight evidence fitting under evidence types (2) and (3), which are otherwise de novo approaches to pattern finding. Equally importantly, this approach can help detect events such as inversions, for which there is otherwise no automated help available, and which can therefore be detected only by comparing sequences. For example, in the *rps*12 gene of a number of plant chloroplasts exon 1 is inverted with respect to exons 2 and 3 (e.g., *Eucalyptus*, *Lotus*, *Oenothera*, *Vitis*). Unfortunately, while this inversion is explicitly noted in the original database sequences of *Lotus* and *Oenothera*, it is not noted for either *Eucalyptus* or *Vitis*. Thus, in the alignment of Jansen et al. (2006) positions 1–114 of the *Eucalyptus* sequence have not been reverse-complemented, while the other three sequences have been. Manually grouping sequences is probably the most common approach to adjusting alignments “by eye”, which is often an undescribed feature of published alignments.

Indeed, the importance of aligning sequences in their taxonomic groups cannot be over-emphasized (see Fig. 5). For example, there is an alignment in Release 17.0 of the Pfam database (Finn et al. 2006) of the homologous chains of the NADH-Ubiquinone/plastoquinone (complex I) (database entry PF00361). If the sequences in this alignment are re-arranged into their two component chains (i.e., mitochondrial *nad*4 and *nad*5) and then into their current taxonomic groups, many misaligned motifs become immediately obvious. However, we must be careful that this sort of procedure is not circular, since one of the most common purposes of a phylogenetic analysis is to test the current taxonomy. My comments above on the independence of proposing and testing hypotheses thus apply here, as well; this procedure works best if used as reciprocal illumination (Ochoterena this volume).

The structure of the encoded products (evidence type 6) has long been considered to be important for sequence alignment (reviewed by Morrison 2006). Indeed, evidence type (1) above is actually the first-order structure, while the zero-order structure (residue composition) has repeatedly been reported to create confounding biases in phylogenetic analyses (Jermiin et al. 2004). The second-order (planar) and third-order (three-dimensional) structures can provide evidence of where evolutionary events are most/least likely to occur, as certain structures place constraints on events, so that different events occur with different frequencies in different structural regions (e.g., due to molecular interactions, or to characteristics such as variable compositional bias).

For RNA-coding sequences, the most important second-order structures are the double-stranded stems. Here the alignment procedure needs to account for doublets, so that the alignment is consistent for both halves of the stem (e.g., Fig. 4). There have been a number of attempts to provide computer programs that correctly align RNA sequences based on their structures (see Morrison 2006), but in my experience these programs fail when confronted with length-variable stems and the absence of predicted stems. Both of these are very common situations in rRNA sequences, for example, especially as sequence similarity decreases, and so this is an active area of research (e.g., Kiryu et al. 2007; Torarinsson et al. 2007; Xu et al. 2007). However, being able to align RNA structures is a different thing from being able to align their primary sequences, and there may be little ability to discern homologous nucleotide positions (e.g., the abbreviated mitochondrial rRNA genes of *Xiphinema americanum* compared to those of other nematodes; He et al. 2005).

Codons make up the important second-order structural feature of protein-coding sequences. They can be accommodated in a sequence alignment by first translating the nucleotides to amino acids and then aligning the amino acids. This can be very successful, because proteins have strong statistical signals in their primary sequence that can be detected at the amino-acid level, such as codon bias (except when there are frame-shifts). Figure 6 shows an example of such an alignment, where the average pairwise sequence identity is <65% and yet the alignment is straightforward once the codon structure is taken into account. In such a structure-based alignment, the nucleotides are aligned against the amino acids within each sequence and then the amino acids are aligned against each other between sequences.

The third-order structures, such as tertiary interactions for RNA-coding sequences and helices/sheets for protein-coding sequences, can also provide evidence. This is shown in the example in Fig. 6, where the protein structures make it clear that all of the length variation is adjacent to the 5′ end of the domain. There is also some variability in pairwise identity between the structural regions of the figure: 62.1% for the helices, 61.6% for the sheets and 65.6% for the remainder of the domain, with 53.6% for the flanking regions. This indicates that there are different evolutionary

```
Synechococcus sp. PCC6301 (1rbl)                                               M  P  K  T  Q  S  A  A  G  Y  K  A  G  V  K  D  Y  -   K  L  T  Y  Y  T  P  D  Y  T  P
Chlamydomonas reinhardtii (1gk8)                                      M  V  P  Q  T  E  T  K  A  G  A  G  F  K  A  G  V  K  D  Y  -   R  L  T  Y  Y  T  P  D  Y  V  V
Alcaligenes eutrophus (1bxn)                                    M  N  A  P  E  S  V  Q  A  K  P  R  K  R  Y  D  A  G  V  M  K  Y  -   K  M  G  Y  W  D  G  D  Y  V  P
Galdieria partita (1bwv)           M  S  Q  S  I  E  E  K  S  V  Q  E  R  T  R  I  K  N  S  R  Y  E  S  G  V  I  P  Y  A   K  M  G  Y  W  N  P  D  Y  Q  V
Structure                          -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -   H  H  H  H  S  -  -  -  -  -  -
Prochlorothrix hollandica         ---------------------------------------ATGGCAGTACAGACCAAAGGCTATCAGGCCGGTGTAAAAGACTAC---|CGCCTGACCTACTACACCCCCGAATACACCCCC
Nostoc sp. PCC7120                ---------------------ATGTCTTACGCTCAAACGAAGACTCAGACAAAATCTGGGTATAAAGCCGGGGTTCAAGATTAC---|AGACTAACTTATTACACACCTGATTACACACCT
Cyanophora paradoxa               ------------------------ATGTCATCACAAGCTAGAACTCAGACAAGAGCAGGCTTCAAAGCAGGTGTAAAAGATTAT---|CGTTTAACTTATTATACTCCTGAATATACTCCA
Abies magnifica                   ------------------------ATGTCACCAAAAACAGAAACTAAAGCTAGTGTCGGATTCAAAGCTGGTGTTAAAGATTAC---|AGATTAACTTATTATACTCCTGAATATCAGACC
Astasia longa                     ------------------------ATGCCAGCTCAAACTGAAAGTAAAACTACCACTGAATTTAAAGCAGGTGTAAAAGATTAT---|AAAATTACTTATTATACTCCTGATTATAAAGTT
Thiobacillus ferrooxidans         ---------------------------------------------ATGGCCGTAAAAACGTATGAGGCCGGTGTAAAGGACTAC---|CGGCAGACCTATTGGGCCCCCGAGTACGTGCCT
Chromatium vinosum (rbcL1)        ------------------------------------------------ATGGCTAAGACGTACAGCGCGGGCGTGAAAGAGTAC---|CGCGAGACCTATTGGATGCCCAACTACACTCCG
Chromatium vinosum (rbcL2)        ---------------------------------------------ATGAGCACGAAGACGTATGACGCCGGCGTCAAAGATTAC---|GCCCTGACCTACTGGACCCCGGACTATGTGCCC
Cyanidium caldarium               ATGTCTCAATCACTAGAAGAAAAATCTGTTCAAGAACGTACAAGAATTAAAAATAGTAGATATGAGTCAGGAGTAATACCATATGCT|AAAATGGGATATTGGAATCCTGATTATCAAGTA
Antithamnion sp.                  ---------------ATGTCTAACTCTGTAGAAGAACGGACAAGAATTAAAAATGAGCGTTATGAGTCTGGTGTAATACCTTACGCT|AAAATGGGATATTGGGATCCTGAATATGCAGTT
Ectocarpus siliculosus            ---------------ATGCCTGAAGATGTGCAAAATCGAACTCGAATAAAAAGTGAACGTTATGAATCAGGTGTTATCCCATATGCC|AAAATGGGATATTGGGATGCTGATTATAAGCTT
Cylindrotheca sp. strainN1        ---------------ATGTCTCAATCTGTATCAGAACGGACTCGAATCAAAAGTGACCGTTACGAATCTGGTGTAATCCCTTACGCT|AAAATGGGTTACTGGGATGCTTCATATGCAGTA
Guillardia theta                  ---------------ATGTCTCAATCCGTAGAATCACGGACTCGAATCAAAAACGAACGTTACGAATCAGGTGTTATCCCTTACGCA|AAAATGGGATACTGGGATGCTGATTACGTTATC
Rhodobacter sphaeroides           ---------------------ATGGATACCAAGACCACCGAGATCAAGGGCAAGGAACGCTACAAGGCCGGCGTCCTGAAATATGCC|CAGATGGGTTACTGGGACGGCGATTACGTCCCC
Ralstonia eutropha                ------------------ATGAACGCACCTGAAACGATCCAAGCCAAGCCGCGCAAGCGCTATGACGCCGGCGTGATGAAGTACAAG|GAAATGGGGTACTGGGACGGCGACTACGTGCCC

Synechococcus sp. PCC6301          K  D  T  D  L  L  A  A  F  R  F  S  P  Q  P  G  V  P  A  D  E  A  G  A  A  I  A  A  E  S  S  T  G  T  W  T  T  V  W  T
Chlamydomonas reinhardtii          R  D  T  D  I  L  A  A  F  R  M  T  P  Q  P  G  V  P  P  E  E  C  G  A  A  V  A  A  E  S  S  T  G  T  W  T  T  V  W  T
Alcaligenes eutrophus              K  D  T  D  L  L  A  L  F  R  I  T  P  Q  D  G  V  D  P  V  E  A  A  A  A  V  A  G  E  S  S  T  A  T  W  T  V  V  W  T
Galdieria partita                  K  D  T  D  V  L  A  L  F  R  V  T  P  Q  P  G  V  D  P  I  E  A  A  A  A  V  A  G  E  S  S  T  A  T  W  T  V  V  W  T
Structure                          -  -  -  -  S  S  S  S  S  S  S  S  S  -  -  -  -  -  H  H  H  H  H  H  H  H  H  H  H  -  -  -  -  -  -  -  -  -  G  G
Prochlorothrix hollandica         AAGGACACAGACCTGCTGGCTTGTTTCCGCATGACCCCCCAGCCCGGTGTCCCCCCCGAAGAAGCTGGTGCTGCGGTTGCTGCTGAATCTTCCACCGGTACCTGGACCACCGTTTGGACT
Nostoc sp. PCC7120                AAAGATACAGATATTCTGGCGGCGTTCCGTGTTACACCCCAGCCCGGAGTTCCCTTTGAGGAAGCGGCTGCGGCTGTAGCGGCTGAGTCTTCTACTGGTACTTGGACGACCGTATGGACA
Cyanophora paradoxa               AAAGAAACTGACATTCTAGCAGCTTTCAGAATGACTCCTCAACCAGGAGTACCTCCTGAAGAGTGTGCAGCAGCAGTAGCAGCAGAATCTTCCACTGGTACTTGGACAACTGTATGGACT
Abies magnifica                   AAAGATACGGATATCTTGGCGGCATTCCGAGTAACTCCTCAACCTGGGGTGCCGCCCGAGGAAGCGGGAGCAGCAGTAGCTGCTGAATCTTCCACCGGTACATGGACCACTGTTTGGACC
Astasia longa                     TCTGATACTGATGTTTTGGCAGCTTTTCGTATGACTCCTCAAACAGGAGTTACTCCAGAAGAATGTGGTGCAGCTGTAGCTGCTGAATCTTCTACAGGTACTTGGACTACTGTATGGACT
Thiobacillus ferrooxidans         TTGGATTCCGATATCCTAGCCTGTTTCAAGATTACGCCGCAACCAGGGGTGGATCGTGAGGAAGCCGCCGCCGCGGTAGCCGCCGAGTCTTCTACCGGCACATGGACCACGGTGTGGACC
Chromatium vinosum (rbcL1)        AAGGACACGGACATCCTGGCCTGCTTCAAGATCACCCCGCAGGCGGGCGTTCCGCGCGAAGAGGCCGCTGCCGCCGTCGCCGCCGAATCCTCGACCGGCACCTGGACCACGGTCTGGACC
Chromatium vinosum (rbcL2)        CTGGATTCCGACTTGCTGGCCTGCTTCAAGGTCACGCCGCAGGCCAAGGTGTCACGCGAGGAGGCCGCCGCCGCCGTCGCCGCCGAGTCCTCGACCGGCACCTGGACCACGGTCTGGTCG
Cyanidium caldarium               AAAGATACAGATGTACTAGCATTATTTAGAGTCACACCTCAACCAGGAGTAGATCCAATAGAAGCTGCTGCTGCAGTTGCTGGAGAATCTTCAACAGCTACATGGACAGTTGTTTGGACA
Antithamnion sp.                  AAAGATACTGATATCTTAGCTCTATTCCGTGTAAGTCCACAACCAGGTGTTGATCCAATTGAAGCATCTGCAGCTGTAGCAGGTGAATCATCTACTGCAACTTGGACTGTTGTATGGACA
Ectocarpus siliculosus            AAAGATACTGATATTCTAGCTCTTTTTCGTATCACTCCACAACCAGGCGTAGATCCAGTTGAAGCTGCAGCAGCTGTTGCCGGTGAATCTTCTACTGCAACATGGACAGTGGTTTGGACA
Cylindrotheca sp. strainN1        AAAACTACTGATGTTTTAGCTTTATTCCGTATTACACCACAACCTGGTGTAGATCCAGTAGAAGCTGCTGCGGCAGTAGCGGGTGAATCTTCTACAGCAACCTGGACAGTTGTATGGACT
Guillardia theta                  AAAGATACTGACGTACTTGCAATGTTCCGTATGACACCTCAAAAAGGTGTTGACCCAGTTGAATGTGCAGCCGCAATCGCAGGTGAATCTTCAACAGCAACTTGGACAGTTGTATGGACA
Rhodobacter sphaeroides           AAGGACACCGACGTGCTGGCGCTGTTCCGCATCACACCGCAGGAGGGGGTGGATCCGGTCGAGGCGGCCGCCGCCGTGGCCGGCGAAAGCTCGACTGCGACCTGGACCGTGGTCTGGACC
Ralstonia eutropha                AAGGACACCGATGTGCTGGCGCTGTTCCGCATCACGCCGCAGGACGGCGTCGATCCGGTCGAAGCCGCCGCGGCGGTCGCGGGCGAATCGTCGACGGCTACCTGGACGGTGGTGTGGACC

Synechococcus sp. PCC6301          D  L  L  T  D  M  D  R  Y  K  G  K  C  Y  H  I  E  P  V  A  G  E  E  N  S  Y  F  A  F  I  A  Y  P  L  D  L  F  E  E  G
Chlamydomonas reinhardtii          D  G  L  T  S  L  D  R  Y  K  G  R  C  Y  D  I  E  P  V  P  G  E  D  N  Q  Y  I  A  Y  V  A  Y  -  I  D  L  F  E  E  G
Alcaligenes eutrophus              D  R  L  T  A  C  D  M  Y  R  A  K  A  Y  R  V  D  P  V  P  N  N  P  E  Q  F  F  C  Y  V  A  Y  D  L  S  L  F  E  E  G
Galdieria partita                  D  L  L  T  A  A  D  L  Y  R  A  K  A  Y  K  V  D  Q  V  P  N  N  P  E  Q  Y  F  A  Y  I  A  Y  E  L  D  L  F  E  E  G
Structure                          G  G  G  -  -  H  H  H  H  -  -  S  S  S  S  S  S  S  -  -  -  -  -  -  -  S  S  S  S  S  S  S  -  G  G  G  -  -  -  -
Prochlorothrix hollandica         GACCTTCTGACCGACCTCGATCGCTACAAAGGTCGTTGCTATGAAGTGGAGCCAGTGCCGGGTGAAGACAACCAGTACTTCTGCTTTGTGGCCTATCCCTTGGACCTTTTTGAAGAAGGT
Nostoc sp. PCC7120                GACCTGTTAACCGATCTAGATCGTTACAAAGGTCGTTGCTACGATATCGAACCAGTTCCCGGCGAAGACAACCAATTCATTGCCTACATCGCTTATCCTTTGGATCTGTTTGAAGAAGGC
Cyanophora paradoxa               GATGGTTTAACCAGTCTTGACCGTTACAAAGGTAGAAGCTATGGTTTCGAACCAGTTCCTGGTGAAGAAAACCAATACATTTGTTACGTAGCATATCCTTTAGATTTATTTGAAGAAGGT
Abies magnifica                   GATGGACTTACCAGTCTTGATCGTTACAAAGGGCGATGCTATGACATCGAGCCCGTTCCTGGAGAGGAAAGTCAATTTATTGCCTATGTAGCTTACCCCTTAGACCTTTTCGAAGAAGGT
Astasia longa                     GATGGTTTAACTCAACTTGATAGATATAAAGGTCGTTGTTATGATATTGAGCCAGTATATGGTGAATGTGATCAATATATAGCTTATATAGCTTATCCTATTGAACTTTTTGAAGAAGGT
Thiobacillus ferrooxidans         GATTTGTTGACCGACATGGATTACTACAAAGGCCGCGCCTATCGCATTGAAGATGTGCCGGGCGACGACACCTGCTTTTACGCGTTCGTTGCCTATCCCATAGATTTGTTTGAAGAGGGG
Chromatium vinosum (rbcL1)        GACCTGCTGACCGACCTGGATTACTACAAGGGCCGCGCCTACGCCATCGAGGACGTGCCGGGCGACGACACCTGCTTCTATGCCTTCATCGCCTATCCGATCGATCTGTTCGAGGAAGGC
Chromatium vinosum (rbcL2)        GATCTGCTGACCGATCTCGATTACTACAAGGGTCGCGCCTACCGGATCGAGGACGTGCCGGGCGACAAGGAGAGCTTCTATGCCTTCATCGCCTATCCGCTCGATCTGTTCGAGGAAGGC
Cyanidium caldarium               GATTTATTAACAGCAGCAGACTTATATAGAGCAAAAGCTTATAAAGTAGATCAAGTTCCTAATAATCCAGAACAATATTTTGCATATATTGCATATGAACTAGATTTATTTGAAGAAGGA
Antithamnion sp.                  GATTTATTAACTGCATGTGATTTATATAGAGCAAAAGCTTACAAAGTAGATGCTGTACCTAACACATCTGATCAATACTTCGCATATATTGCATATGATATCGATTTATTTGAAGAAGGT
Ectocarpus siliculosus            GATTTATTAACTGCTTGTGATATCTATAGAGCAAAAGCATATAGAGTAGATCCAGTACCAGGCACAAGTGATCAATTCTTTGCTTATATAGCTTACGAATGTGATTTATTTGAAGAAGGT
Cylindrotheca sp. strainN1        GATTTATTAACAGCTTGTGACCGTTACCGTGCTAAGGCTTACCGAGTAGATCCAGTTCCAAATGCAGCAGATCAATACTTTGCTTTCATCGCATACGAATGTGATTTATTTGAAGAAGGT
Guillardia theta                  GACCTTCTAACAGCATGTGATCTTTACCGTGCAAAAGCATACCGTGTTGACCCAGTTCCTGGTGCAACTGATCAATACTTTGCTTACATCGCATACGAATTAGACCTATTTGAAGAAGGT
Rhodobacter sphaeroides           GACCGGCTGACCGCCTGCGACAGCTACCGCGCCAAGGCCTACCGGGTGGAGCCGGTGCCGGGCACGCCGGGGCAGTATTTCTGCTATGTGGCCTATGACCTGATCCTGTTCGAGGAAGGC
Ralstonia eutropha                GACCGCCTGACTGCCTGCGACATGTACCGCGCCAAGGCCTACCGGGTCGACCCGGTGCCCAACAACCCCGAGCAGTTCTTCTGCTACGTGGCCTATGACCTGTCGCTGTTCGAGGAAGGC

Synechococcus sp. PCC6301          S  V  T  N  I  L  T  S  I  V  G  N  V  F  G  F  K  A  I  R  S  L  R  L  E  D  I  R  F  P  V  A  L  V  K   T  F  Q  G  P
Chlamydomonas reinhardtii          S  V  T  N  M  F  T  S  I  V  G  N  V  F  G  F  K  A  L  R  A  L  R  L  E  D  L  R  I  P  P  A  Y  V  K   T  F  V  G  P
Alcaligenes eutrophus              S  I  A  N  L  T  A  S  I  I  G  N  V  F  S  F  K  P  I  K  A  A  R  L  E  D  M  R  F  P  V  A  Y  V  K   T  F  A  G  P
Galdieria partita                  S  I  A  N  L  T  A  S  I  I  G  N  V  F  G  F  K  A  V  K  A  L  R  L  E  D  M  R  L  P  L  A  Y  L  K   T  F  E  G  P
Structure                          -  H  H  H  H  H  H  H  H  H  -  -  G  G  G  -  -  -  S  S  S  S  S  S  S  S  S  S  -  -  H  H  H  H  -   -  -  -  -  -
Prochlorothrix hollandica         TCTGTCACCAACATTCTGACCTCCATCGTCGGTAACGTGTTTGGCTTCAAAGCCCTGCGTGCCCTGCGTTTGGAAGATATCCGCTTCCCCATCGCCCTGGTCAAG|ACCTTCCAAGGTCCT
Nostoc sp. PCC7120                TCCATCACCAACGTTTTGACCTCAATTGTAGGTAACGTATTTGGTTTTAAAGCATTACGCGCATTGCGTTTGGAAGACATTCGCTTTCCTGTTGCTTACATCAAG|ACCTTCCAAGGCCCT
Cyanophora paradoxa               TCTGTTACTAACATGTTAACTTCCATCGTAGGTAACGTATTTGGTTTCAAAGCATTACGTGCATTACGTTTAGAAGATTTACGTATTCCAGTTGGTTATTCCAAA|ACTTTCCAAGGTCCT
Abies magnifica                   TCTGTTACTAACTTGTTCACTTCCATTGTAGGTAATGTATTTGGATTCAAGGCCCTACGGGCTCTACGTTTGGAAGATTTGCGGATTCCCCCTGCTTATTCCAAA|ACTTTTCAAGGTCCA
Astasia longa                     TCTGTTACTAATCTTTTGACAAGTATTGTAGGAAATGTTTTTGGTTTTAAAGCTTTAAGAGCTCTTCGTTTAGAAGATTTACGTATTCCAGCTTGTTATTTAAAA|AGTTTTTGGGGACCA
Thiobacillus ferrooxidans         TCCGTGGTTAACGTCTTCACCTCCCTGGTCGGTAACGTGTTCGGCTTCAAGGCCGTGCGTGCGCTGCGCCTGGAGGATGTGCGATTCCCGCTCGCCTATGTCAAG|ACTTGCAATGGACCG
Chromatium vinosum (rbcL1)        TCGGTCGTCAACGTCTTCACCTCGCTCGTCGGCAACGTGTTCGGCTTCAAGGCCGTGCGCGCCCTGCGTCTGGAGGACGTCCGCTTCCCGATCGCCTATGTCATG|ACCTGCAACGGCCCG
Chromatium vinosum (rbcL2)        TCGATCGTCAACGTGCTCACCTCGCTGGTGGGCAACGTGTTCGGCTTCAAGGCGGTGCGCGCCCTGCGTCTGGAAGACATCCGCTTCCCGCTGCACTATGTGAAG|ACCTGCGGCGGCCCG
Cyanidium caldarium               TCAATTGCGAACTTAACTGCTTCTATAATTGGTAATGTATTTGGATTTAAAGCAGTTAAAGCTCTAAGACTTGAGGATATGAGATTGCCATTTGCATATATTAAA|ACTTTCCAAGGACCT
Antithamnion sp.                  TCAATTGCAAACTTAACTGCTTCAATTATCGGTAACGTATTTGGTTTCAAAGCAATAAAAGCTTTAAGATTAGAAGATATGCGTATTCCTGTAGCTTACCTAAAA|ACAATCCAAGGTCCA
Ectocarpus siliculosus            TCTCTTGCTAATTTAACAGCATCTATTATTGGTAACGTTTTCGGATTCAAAGCTGTTAAAGCGTTACGTTTAGAAGATATGAGAATCCCTTATGCTTACTTAAAA|ACTTTCCAAGGCCCA
Cylindrotheca sp. strainN1        TCGTTAGCTAACTTAACAGCATCGATTATTGGTAATGTTTTTGGCTTTAAAGCTGTAGCAGCTTTACGTTTAGAAGATATGCGTATTCCTCACTCATACTTAAAA|ACATTTCAAGGTCCT
Guillardia theta                  TCTTTAGCTAACTTAACAGCATCAATTATTGGTAACGTATTCGGATTCAAAGCGGTAAATGCTTTAAGATTAGAAGATATGCGTCTTCCAATTGCATACCTAAAA|ACTTTCCAAGGTCCA
Rhodobacter sphaeroides           TCGATCGCCAACCTCACCGCCTCGATCATCGGCAACGTCTTCAGCTTCAAGCCGCTGAAGGCCGCCCGGCTCGAGGACATGCGCTTCCCCGTGGCCTATGTGAAG|ACCTACAAGGGCCCG
Ralstonia eutropha                TCGATCGCCAACCTGACCGCCTCGATCATCGGCAACGTGTTCAGCTTCAAGCCGATCAAGGCGGCGCGCCTGGAAGACATGCGCTTCCCGGTCGCCTACGTGAAG|ACCTTCGCCGGCCCG
```

Fig. 6 Aligned N-terminal domain and flanking regions of the chloroplast *rbc*L gene of 15 nucleotide sequences from selected bacteria and plastid-bearing species. The *first four lines* show the amino-acid sequences of some molecules that have had their structure experimentally determined (the PDB database code is shown after the species name). The fifth line shows the consensus secondary structure from those sequences (for illustrations see Kellogg and Juliano 1997): *H* alpha-helix, *G* 3(10) helix, *S* H-bonded beta-strand or isolated beta-bridge; – random coil, bend or H-bonded turn. The *vertical bars* delimit the domain. The sequences are aligned within the domain and for the 3′ suffix, but only the nine amino acids immediately 5′ to the domain can be unambiguously aligned (the remainder have been right-aligned for convenience of presentation). The data are taken partly from each of the Homstrad (Stebbings and Mizuguchi 2004), PANDIT (Whelan et al. 2006) and Pfam (Finn et al. 2006) databases (Version 17.0, domain family PF02788 in the latter two). In a structure-based alignment such as this, the sequences are aligned within the domains, the nucleotides being aligned against the amino acids within each sequence and the amino acids aligned against each other between sequences. The sequences are often not aligned for the flanking regions, the amino acids simply being moved to one end or the other of the region. However, in this example the prefix flank can be aligned for nine amino acids adjacent to the domain; and all seven amino acids of the suffix flank, which joins the two *rbc*L domains, can be aligned (only five of the latter amino acids are shown here)

constraints on the various regions, which can affect the search for positional homology if it is based solely on sequence similarity. There has been much research on the development of programs for the alignment of amino-acid sequences taking into account third-order structures (see Morrison 2006). Fourth-order structure (inter-molecular complexing) is rarely useful for phylogenetic purposes, mainly because we know very little about it in many cases.

Non-coding sequences such as group I & II introns and internal transcribed spacers (and also many inter-genic spacers) have structures that are similar to those of rRNAs and tRNAs (Damberger and Gutell 1994; Kelchner 2000, 2002; Schultz et al. 2005). For this combined grouping, the sequences are usually a mosaic of four distinct types of region: (1) single-stranded regions that are highly conserved, both with respect to length and composition (these represent the functional sites), (2) non-conserved single-stranded regions, which can show great variability in length and composition (these allow the fixed parts of the RNA molecule to fold into place), (3) conserved stem (i.e., paired) regions (which hold the functional sites into the necessary position), and (4) length-variable stem regions (which have no common function).

The type-(1) regions are easily aligned using similarity. The type-(3) regions can be aligned in an objective and repeatable manner once the nucleotide-pairing rules have been taken into account, as there are particular motifs that are associated with the paired nucleotides of stems. For example, inverted repeats are associated with stems, as shown in Fig. 1 and 2. Furthermore, inversions are often delimited by inverted repeats (Kelchner and Wendel 1996; Graham et al. 2000; Quandt et al. 2003), where the two parts of the repeat form the stem and the inversion forms the terminal loop (see Fig. 3). Most of the reliable phylogenetic information is likely to be in the type-(3) regions (Morrison 2006).

For the type-(2) and type-(4) regions it may be impossible to find any unequivocal evidence of homology between sequences (except at the base on the type-(4) stems). Type-(2) regions are often referred to as hotspots (in non-coding regions; Kelchner 2000) or hypervariable regions (in RNA-coding regions; Gillespie 2004). They are the major source of length variation in most RNA-coding and non-coding sequences (Borsch et al. 2003; Quandt et al. 2004; Korotkova et al. this volume), and are frequently associated with tandem repeats and other repetitive elements. However, type-(4) stems are also well-known in many introns (e.g., Quandt et al. 2004) and in rRNA sequences (where they are usually called expansion segments; Gillespie 2004). Both types of region are regularly excluded from tree-building analyses, because the a priori hypotheses of homology are unclear in the primary sequence.

One practical problem with the use of information from encoded products is that in order to apply different models to different sequence regions we need to know the gene boundaries, in order to distinguish protein- and RNA-coding genes from each other and from non-coding regions, and to distinguish structural regions. However, we usually need to use the multiple alignment in order to locate these boundaries, in the sense that it is the common pattern across sequences that indicates where the boundaries are located. Sometimes boundaries can be identified in a single sequence even in the absence of independent evidence, but more often this approach leads to incorrect decisions, and the correct decisions come solely from comparing several sequences. An example is shown in Fig. 7, where the annotated gene boundaries do not even match the structural regions within the genes.

The final piece of evidence that I will mention here (evidence type 7) is simply the failure of database searches to detect close matches. Close matches are usually

```
Structure                                   <<<<<<<<<<     >>>>>>>> >>       18S]  [ITS1          ITS1]  [5.8S << <<<<<<<<<<     <<<
(AJ421838) Rhinonyssus tringae              TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [GAGG...ATGATCCATT]  [A AAGAC|TCAATGTGGGGATCAC...
(AJ421837) Tinaminyssus streptopelioides    TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [GTGA...AAAAGT-TTG]  [C AAGAC|TCAATATGGAGGATCAC...
(AJ421832) Tinaminyssus minisetosum         TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [GTGA...ATGAGT-TAG]  [C AAGAC|TCAATATGGAGGATCAC...
(AJ421829) Tinaminyssus columbae            TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [ATGA...ATGAGT-TAG]  [C AAGAC|TCAATATGGAGGATCAC...
(AJ421835) Sternostoma strandtmanni         TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [CTGA...A--ACTATTG]  [C AAGAC|TCAATATGGGGATCAC...
(AJ421834) Sternostoma boydi                TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [CTGA...A--ACTATTG]  [C AAGAC|TCAATATGGGGATCAC...
(AJ421836) Sternostoma turdi                TTTCCGTAGGTGA ACCTGCGGA|AGGATCATTA]  [ATGA...A-GATAATGT]  [C AAGAC|TCAATATGGGGATCAC...
(AF544014) Tropilaelaps koenigerum          ????CGTAG-TGA|ACCTGCGGA AGGATCATTA]  [CTGT...A-GAACGCAT]  [C|AAGAC TCAATATGGGGATCAC...
(AF544013) Tropilaelaps clareae             ????CGTAG-TGA|ACCTGCGGA AGGATCATTA]  [CTGT...AAGAACGCAT]  [C|AGGAC TCAATATGGGGATCAC...

Structure   <<<<<<<<<<<< <<< ]  [ITS2                                                  ITS2]  [28S>>> >>>>>> >>>>>> ^^    <<<<<<
(AJ421838)  CAGTTGTTTCAGCATAT]  [GAGGAGTAC--ATTACGTAGTTACT ATTGCTG|GGATGTAATAG...GCTGATTGTT]  [GTGTAT CTGAAA-CAAGTGTGATGA|CCCCCT
(AJ421837)  CATTTGTTTCAGTATAA]  [AAAAGTACT--AATGCGTAGTTACT ATTGCTG|CAACGCAATAG...ATGAGTTAAC]  [GTGTAT CTGAAATCAAGTGTGATTA|CCCCCT
(AJ421832)  CACTTGTTTCAGTATAT]  [TGACCTAGC--AATACGTAGTTACT ATTGCTG|CAATGCAATAG...TGTAATGGAC]  [GTGTAT CTGAACTCAAGTGTGAGAA|CCCCCT
(AJ421829)  CACTTGTTTCAGTATAT]  [TGACATAGA--AATACGTAGTTACT ATTGCTG|CGATGCAATAG...TGTAATGGAC]  [GTGTAT CTGAACTCAAGTGTGAGAA|CCCCCT
(AJ421835)  CATTTGTTTCAGTATAT]  [AAAAGATGTCCAATACGTAATTGCT GTTGCTG|GAAAGCAATGG...TTGGTGTAGC]  [GTGTAT CTGAAATCAAGTGTGATGA|CCCCCT
(AJ421834)  CATTTGTTTCAGTATAT]  [AAAAGATGTCCAATACGTAATTGCT GTTGCTG|GAAAGCAATGG...TTGGTGTAGC]  [GTGTAT CTGAAATCAAGTGTGATGA|CCCCCT
(AJ421836)  CATTTGATTCAGTATAT]  [AGACTGTAG--AATACGTAGTTACT ATTGCTG|GAATGCAATGG...TAGATATGAC]  [GTGTAT CTGAAATCAAGTGTGTTGA|CCCCCT
(AF544014)  CACTTGTTTCAGTATAT]  [AACTCGTCG--TATAAGTACTGACT|ATTGCCG TTATGCAATGG...TCAAGTTGAC]  [GTGTAT|CTGAAATCAAGTGTGA??? ??????
(AF544013)  CACTTGTTTCAGTATAT]  [AACTCGTAG--TATATGTACTTACT|ATTGCCG T-ACGCAATGG...TCAAGTTGAC]  [GTGTAT|CTGAAATCAAGTGTGA??? ??????
```

Fig. 7 DNA sequences at the four gene boundaries of the ribosomal 18S–ITS1–5.8S–ITS2–28S region of nine species of mites (Acari; Mesostigmata) from the DDBJ–EMBL–Genbank database (accession numbers in *brackets*). The *vertical bars* show the gene boundaries as annotated in the database, whereas the *square brackets* show the true gene boundaries (based on the structure of the RNA molecules). Above each sequence, the angle brackets show the paired nucleotides in the secondary and tertiary structure of the rRNA molecules, which also do not agree with the annotated gene boundaries. The ellipses indicate parts of the sequence that are not shown, and the question-marks are undetermined nucleotides

interpreted as indicating homology between the sequences. However, the important point is not the overall match between the sequences but the pattern of matching. Close matches in one part of the sequence but not in others can indicate that the non-matching regions are not homologous, which is obviously an important consideration in a phylogenetic alignment. For instance, this process can identify regions of a sequence that should not be aligned against the other sequences, but instead should be part of a staggered alignment (Cammarano et al. 1999 present a detailed example).

On a slightly different topic, homology of whole sequences (e.g., gene orthology) is not an easy thing to assess, but is tied up with the concepts of similarity and motifs discussed above (Dessimoz et al. 2005). Homology of the sequences requires an assessment of: (1) global similarity, (2) local similarity, and (3) conservation of motifs. That is, sequence homology is usually first detected using a database search (e.g., a significant match to a database consensus sequence), but it must be confirmed by some acceptable degree of global and local similarity to the other sequences, and must share some conserved motifs (presumably functional residues).

An example is shown in Fig. 8 for a protein-coding sequence. Here, there are no universally conserved alignment positions among the members of the domain family, the best conservation being 50 out of 55 members, and the longest motif shared pairwise among the sequences consists of only seven contiguous identical amino acids (out of the 78–100 amino acids in the domain). Moreover, the longest perfect match between the query sequence and the database consensus (ELGL) does not occur in any of the individual family sequences, although the second-longest (NLR) does match some of the sequences. However, the range of genetic distances from the query sequence to the other sequences is well within the range among the other sequences (global similarity), it shares a motif of five amino acids with one family member and motifs of four amino acids with 11 other members (local similarity), and it shares eight out of nine of the best-conserved residues (motif similarity). Thus, the evidence for homology is as good as it is for any of the known family members.

Current computerized algorithms

The question obviously now arises as to whether any of the currently available computer programs implement the above ideas for sequence alignment and, if so, to what extent they succeed in producing useful phylogenetic alignments. I will start this discussion by pointing out that similarity-based progressive-alignment procedures cannot be expected to work in a phylogenetic context.

The simplest way to make this point is to consider an explicit example where the phylogenetic history of the sequences is known. Sanson et al. (2002) experimentally produced a perfectly balanced molecular phylogeny of *Trypanosoma cruzi* (Kinetoplastida), recording the SSU rRNA of the organisms for the 16 terminal sequences as well as for the 15 ancestors (including the common ancestor). The Clustal program (the most commonly used alignment program in phylogenetic studies) fails to produce the correct alignment if only the 16 terminals are used, but it can do much better if all 31 sequences are included. That is, knowledge of the ancestral sequences provides valuable information that is not available in the terminal sequences. The improvement occurs in those positions where there are closely adjacent (but independent) deletions (as shown for one region in Fig. 4), which is the primary problem also identified by Golubchik et al. (2007). Without knowledge of the ancestors, Clustal incorrectly aligns those residues that are associated with these independent deletions. However, when the ancestors are included, Clustal proceeds by progressively aligning one of each descendant sequence-pair against its immediate ancestor, and then aligning the other descendant to that paired alignment. By doing this it correctly identifies the independent deletions. Unfortunately, it does not produce a completely correct alignment because it gets the guide tree

```
           1        10        20        30        40        50        60        70        80        90        100
ASA1       KLIDYFRDHPSLWNTKNKDYNNRLLRTQ-KLQIIGDELGLT------RKDVYEKYRNLRTTFFREHKRVTRGSCKANALGANGPHLQGYVSRWRHYNHMLFLTIKSS
CDD        RLIELVRERPCLWDRRHPDYRNKEVKRE AWAEIAEELGLS------VEECKKRWKNLRDRYRRELKRLQ-----------NGKSGGGKKSKWKYFERLSFL-----
Modal      RLIELVR?RPCLWDRRHPDYRNREVK???AWEEIAEELGLS??????VEECKKRWKNLRDRYRRELKRVQ???????????????????SKW?YFEELSFLRP?IR
Conserved  --I--------LW-----------------W-------------------------LR--------------------------------W-------FL-----
```

Fig. 8 The MADF domain of the ASA1 amino acid sequence from the study of Ljunggren et al. (2006), aligned against various database sequence summaries. The boldface ASA1 residues match those of the Modal sequence (and usually also the CDD sequence). The CDD sequence is the consensus sequence from the Conserved Domain Database v.2.11 (Marchler-Bauer et al. 2005), with which the ASA1 sequence has significant similarity ($E = 4e^{-14}$); the boldface residues differ from those of the Modal sequence. The Modal sequence shows the most common amino acid in the alignment from the SMART v4.0 protein-domain database (Letunic et al. 2004), while the Conserved sequence shows those amino acids occurring in >70% of the sequences in the SMART alignment. The underlined residues are the conserved blocks as identified from the SMART alignment by the BlockMaker program (Henikoff et al. 1995), using the Motif algorithm of Smith et al. (1990) and the Gibbs Sampling strategy of Lawrence et al. (1993)

wrong. The correct unrooted tree is produced but the mid-point rooting places the root in the wrong place, and so the final steps of the progressive alignment are incorrect.

It is thus clear that even in simple cases like this (16 sequences, aligned length = 2,236 bp, identity = 98.4%) we cannot expect contemporary sequences to have enough historical traces for us to be able to use similarity alone to construct alignments correctly. We therefore need to evaluate the relative effectiveness (or ineffectiveness) our current alignment procedures. Unfortunately, previous evaluations of automated alignment procedures have concentrated on structure-based alignments as their gold standard (recently reviewed by Wilm et al. 2006 for RNA-coding sequences, and Pei and Grishin 2007 for protein-coding sequences), rather than focusing on phylogenetic alignments. Furthermore, none of the current programs that are commonly used in phylogenetic analyses explicitly implement many of the above ideas, except for sequence similarity and sometimes information from second-order structure (reviewed by Morrison 2006). There is thus no a priori reason to expect these programs to succeed other than in the sense of providing a useful heuristic tool. It is therefore worthwhile for me to directly compare some of these programs on a set of multiple-sequence alignments, to see how well they fare as heuristics.

To this end, I manually prepared a set of seven multiple-sequence alignments based on the ideas outlined above, covering a range of sequence types, lengths and identities (Table 1). In each case, I started with the originally provided alignment, and then manually adjusted this in order to create the most parsimonious set of scenarios for the evolutionary events leading to the contemporary sequences. This was thus effectively a refinement procedure, where the preliminary alignment was assessed for plausibility in terms of the evolutionary events; it is discussed in more detail in the next section. Sometimes this procedure resulted in quite large changes to the alignment and sometimes not. In no case do I guarantee to have found the optimal (most parsimonious) alignment, but I expect that my alignments are closer to this goal than were the originals, and that my scenarios are more plausible. Staggered alignments were used in cases of doubtful homology. These alignments are available at: http://hem.fyristorg.com/acacia/alignments.htm.

Table 1 compares seven computerized algorithms in terms of their ability to reproduce my alignments. Most of these algorithms are described by Morrison (2006) as exemplifying the range of currently available techniques used in phylogenetics. I made no attempt to optimize the parameter values for any of the computer programs, but I

Table 1 Success of several computer programs for the phylogenetic alignment of seven data sets with varying length and similarity of DNA sequences

Data set name[a]	Coding type	Number of sequences	Aligned length (bp)	Average pairwise identity (%)	Alignment success (%)[b]						
					ClustalW	MAFFT	ProbCons	Prank	AliFritz	POY-p	POY-ml
FN1	Protein	18	144	44	68.3	79.2	71.5	48.1	39.1	61.6	29.1
rbcL	Protein	15	534	63	97.9	99.7	99.3	98.4	96.9	94.7	72.4
HSP70	Protein	22	2,048	87	98.7	99.8	98.4	96.1	95.9	95.5	95.0
PRO	tRNA	22	74	60	84.2	89.7	94.2	77.7	71.7	81.0	76.8
Isospora	rRNA	21	1,993	95	96.0	96.5	97.1	96.0	94.4	95.3	95.4
ITS2	Transcribed spacer	22	182	78	87.8	92.1	91.4	87.4	80.0	77.8	74.8
petD	Intron	21	1,184	91	94.8	96.0	97.6	97.0	93.1	92.2	92.9

Success was measured as the pairwise similarity (sum-of-pairs score, as calculated by the Compare program of Do et al. 2005) to a manually constructed phylogenetic sequence alignment

[a] Sequence sources: FN1 (paralogues of the type-I-repeat of the fibronectin gene from species of vertebrates): see Fig. 10; rbcL (large subunit of the chloroplast ribulose-1,5-biphosphate carboxylase gene from species of eukaryotes and bacteria): see Fig. 6; HSP70 (nuclear 70 kDa heat-shock protein gene from isolates of *Cryptosporidium* species (Apicomplexa)): Xiao et al. (2002); PRO (mitochondrial proline transfer-RNA gene from species of nematodes (Nematoda)), J. Höglund et al. (unpublished); Isospora (nuclear small-subunit ribosomal-RNA gene from species of coccidia (Sporozoa)): Morrison et al. (2004); ITS2 (nuclear rRNA internal transcribed spacer 2 from species of mites (Acari)): Morrison (2006); petD (group II intron of the C-terminal subunit of the chloroplast cytochrome b6 (i.e., petD) gene from species of eumagmoliids (Angiospermae)): Löhne and Borsch (2005)

[b] Computer programs: ClustalW v.1.83 with default parameters (Thompson et al. 1994); MAFFT v.5.861 with default parameters (Katoh et al. 2005); ProbCons v.1.11 with default parameters (Do et al. 2005); Prank v.070126 with default parameters (Löytynoja and Goldman 2005); AliFritz v.1.0 with the default parameters and the simulated annealing stopped after 40 NNI (Fleissner et al. 2005), except for HSP70 and Isospora which used 20 and 25 NNI, respectively; POY v.3.0.11 implied alignment based on parsimony (Wheeler 1996) with the default parameters and the search strategy described by Giribet et al. (2001); POY v.3.0.11 implied alignment based on the total-likelihood model of Wheeler (2006) with the default parameters and the search strategy described by Giribet et al. (2001), except for HSP70 and Isospora which used the abbreviated search strategy of Wheeler (2006)

merely accepted the defaults. It is thus possible that improved solutions could be obtained from each program under some other combination of parameter values (Wheeler and Kececioglu 2007). I did, however, perform more extensive searches for the POY program than are implemented by the program's defaults (see below).

The similarity-based alignment algorithms (ClustalW, MAFFT and ProbCons) assume that there is only one sort of event (substitution) and then insert gaps into the alignment in order to minimize the number of substitutions (=maximize among-sequence similarity). They do not have an explicit model of indels, but instead model them as a set of substitutions with variable weights (based on the gap opening and gap extension penalties). This pattern-matching approach works well when there are non-overlapping or non-adjacent indel events (Golubchik et al. 2007). However, the probability of overlapping or nearby indels increases as the evolutionary distance between the sequences increases, and so these methods start to fail (Morrison 2006). In the example here, decreased pairwise identity of the sequences is clearly associated with decreased alignment success for all three sequence types (Table 1). The fact that the similarity alignment procedures do not create staggered alignments (i.e., they over-align) is their ultimate weakness. This is related to the statistical concepts of sensitivity and selectivity (Lambert et al. 2003). In order to increase sensitivity (thus reducing false negatives, consisting of unaligned positions that should be aligned) the similarity programs sacrifice selectivity (thus increasing false positives, consisting of aligned positions that should not be aligned).

However, those algorithms that use iterative refinement of an initial alignment (MAFFT) or use some measure of consistency across multiple sequences (ProbCons) try to take a more global perspective of the alignment than is possible for the methods based on strictly pairwise comparisons (Wheeler and Kececioglu 2007). This can help to identify evolutionary events based on patterns of similarity across multiple sequences. In this set of examples, these two programs are shown to have quite successful heuristic strategies, as either MAFFT or ProbCons has the highest score for each example (Table 1), usually with the other program ranked second. Interestingly, MAFFT was more successful for the protein-coding sequences and ProbCons for the RNA-coding ones, although it would be premature to generalize this observation.

The algorithm used by the Prank program explicitly recognizes the difference between insertions and deletions, and deals with the fact that the subsequent history of the sequences must be different after each of these events. In theory, it should be the program that is most capable of creating staggered alignments. However, the algorithm does not try to reconstruct particular evolutionary events, and it is still restricted to pairwise comparisons of the sequences based on single nucleotides. It thus does not cope well with the example alignments used here (Table 1), a conclusion also reached by Golubchik et al. (2007).

The statistical-alignment algorithms (AliFritz in this example) explicitly add the concept of indels to the concept of substitutions when modeling the sequences. However, the current methods do not try to model evolutionary events, but are based on standard likelihood substitution models with an indel model added. They are thus based on one simple model of substitutions and one simple model of indels. They differ in how the indel model is implemented, either treating each alignment position as a separate indel (the TKF1 model) or having indivisible fragments (the TKF2 model), the latter based on the empirical observation that indels frequently cover multiple positions (Pons and Vogler 2006; Lunter 2007). However, there are multiple phenomena that create apparent substitutions (e.g., substitution, inversion) and apparent indels (e.g., repeats, translocations, deletions, insertions), and the simple likelihood models seem to be unable to cope with the resulting complexities. Consequently, the results are among the worst for my example alignments (Table 1). Dealing with multiple models may not be a straightforward extension of the current algorithms, because the trick is to work out which models to apply to which parts of the sequences.

Direct optimization (as used in POY) is apparently based on an appropriate idea, because it tries to reconstruct ancestral sequences, and so moves closer to the concept of reconstructing ancestral events. Thus, identifying the postulated historical events leading to the alignment is treated as being part of the alignment procedure, rather than being a result of it. However, the method produces trees rather than alignments, and thus derives an alignment directly from a tree, so that it is more accurately described as a "synapomorphy scheme" rather than a true alignment. Furthermore, the basic model treats each nucleotide position as a separate unit, so that events that affect multiple nucleotides are not dealt with in a direct manner (Pons and Vogler 2006). Consequently, for my examples this approach was never particularly successful (Table 1), a conclusion also reached by Ogden and Rosenberg (2007) based on simulated alignments and Kjer et al. (2007) using structure-based alignments. The fixed-states approach to this issue (Wheeler 1999) treats fixed fragments as the basic units when constructing the phylogenetic tree, but these fragments need to be pre-defined, and there can then be no implied alignment. Interestingly, the parsimony criterion consistently did better (or no worse) than the maximum-likelihood criterion (Table 1), which may actually say more about my personal method of alignment than about the relative worth of these two criteria.

The statistical alignment and direct optimization methods rely on search strategies to find their optimal solution, and both AliFritz and POY implement heuristic searches rather than guaranteeing to find the optimum. It can therefore be argued that their poor performance here might be due not to their inappropriate algorithms but to inadequate searching—in both cases the user has to decide how thorough a search should be in terms of time and strategies. To examine this possibility I compared several search strategies for the POY analyses, comparing the default strategy (not very thorough) with those of Wheeler (2006) (more thorough) and Giribet et al. (2001) (very thorough, and not implementable in reasonable time for two of my data sets under likelihood; Table 1). The results show that extra searching always improved the tree score but had no consistent effect on the alignment score (Fig. 9), a conclusion also reached by Ogden and Rosenberg (2007) based on simulated alignments. When the original alignment score was small (e.g., <60%) the extra searching improved the score, but not necessarily otherwise. For the likelihood criterion, the extra searching employed by the Giribet et al. (2001) strategy compared to that of Wheeler (2006) did not necessarily improve the alignment score. Thus, the difference between the reference alignments and the POY alignments is due to the criterion not to the heuristics.

Basically, all of the alignment programs examined here fail to reproduce the semi-manual alignments because they view each alignment horizontally (in its standard orientation) whereas humans look at it vertically. That is, we compare the same block of positions across all of the sequences (a "global" view) while the programs compare sequences pairwise (or in small clusters) along their length. It is those programs that try to go beyond pairwise comparisons at some stage in their algorithm (e.g., refinement for MAFFT and consistency for ProbCons) that are the most successful heuristics for a phylogenetic alignment. However, it is possibly worth pointing out that all of the programs also fail on the *T. cruzi* alignment referred to in a previous section. Indeed, their order of success is: Clustal > Prank > ProbCons > MAFFT > AliFritz > POY. Both POY and AliFritz produce a tree with some unresolved branches and the root in the wrong place.

It thus seems that the most effective current strategy for phylogenetic alignment is to produce an initial alignment via some helpful automated procedure (perhaps using one of the programs assessed here), and then to modify this alignment in response to whatever historical events can be identified. The objective is essentially to annotate the alignment with a plausible and parsimonious scenario of events. This is a tedious procedure, as I can attest.

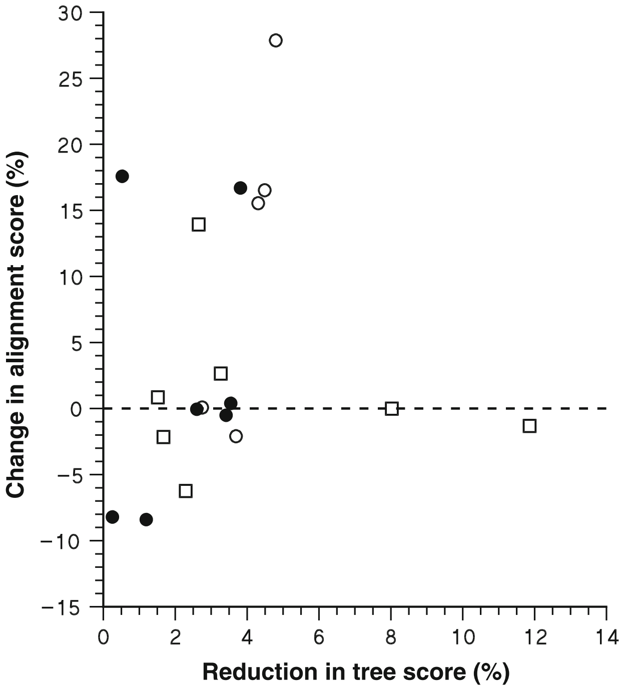

Fig. 9 The effect of performing more extensive searches using the POY computer program, on both the parsimony or likelihood score of the final tree and the alignment success with respect to a phylogenetic alignment, based on the seven data sets described in Table 1. The open symbols compare the default search with the thorough search strategy described by Giribet et al. (2001) for the parsimony (*square symbols*) and likelihood (*circles*) criteria, while the *filled circles* compare the default search with the abbreviated search strategy of Wheeler (2006) for the likelihood criterion. The six analyses with the largest improvement in alignment score all had scores originally <60%, while the three analyses with the largest degradation in alignment score all had scores originally 60–80%

Developing a new algorithm

Aligning events (based on the traces left by those events) can be done manually, by using our biological knowledge. Biological phenomena are rarely optimal in any mathematical sense, because contemporary observations are the result of a series of historical accidents. In this sense, alignment is an inference problem rather than an optimization problem. That is, we infer common ancestors of sequences (based on shared derived character states)—this is the natural way for a biologist to view this particular problem. Optimizing an objective function is thus merely a mathematical convenience, based on the assumption that it should be possible to create a function whose optimal solution is near to the true "biological solution". Inference can be treated mathematically, but there can be biological solutions to biological problems as well as mathematical ones, and so objective and repeatable methods do not necessarily involve mathematics alone (Kjer et al. 2006, 2007).

For example, we can look in the sequences for independent evidence of each of the types of events that can occur, based on known molecular mechanisms, as outlined above. Indeed, this could be done using a rule-based algorithm, where the rules are evaluated sequentially to decide which one applies best at each position in the alignment. The objective would be to minimize the size of a set of explicitly stated evolutionary events, and the rules would relate to evaluating the evidence in relation to which events are possible/likely under which circumstances. This approach can also be used to provide a measure of alignment quality, based on the fit of the data to at least one of the rules. Furthermore, if none of the rules applies particularly well, then an arbitrary convention can be applied as the final rule, which would make the procedure objective and repeatable.

There are several papers that have provided such rules/guidelines (or their associated templates) for non-coding sequences (e.g., Golenberg et al. 1993; Kelchner and Clark 1997; Hoot and Douglas 1998; Graham et al. 2000; Kelchner 2002; Borsch et al. 2003; Löhne and Borsch 2005) and RNA-coding sequences (Kjer et al. 1994; Kjer 1995; Hickson et al. 1996; Kjer 1997; Gillespie 2004). The ultimate rule should always be to use a staggered alignment (i.e., do not align anything) if application of the rules does not lead to a plausible set of hypotheses concerning homology.

We could try to translate this approach into a computerized expert system, although this is never straightforward. Expert systems have not worked all that well in those parts of biology to which they have been applied (e.g., medical diagnosis, taxonomic identification). There is usually some heuristic procedure that does just as good a job and does it much faster (e.g., we now use multi-entry keys for interactive identification, whereas an expert system would be based on a dichotomous key). In general, trying to directly emulate human procedures is not necessarily the best solution to any given problem (e.g., we design moving machines that have wheels rather than legs).

It is the fact that many sequences have repeated and rearranged elements that creates a lot of the problems with the current heuristic algorithms, as these features violate the assumptions on which most of the alignment programs are based. A number of specialist computer programs have been developed with heuristics that accommodate these features in multiple alignments, notably for protein-coding sequences, such as RAlign (Sammeth and Heringa 2006), ABA (Raphael et al. 2004), CombAlign (Wegner et al. 2004) and ProDA (Phuong et al. 2006). Similarly, those recent programs designed to align whole genomes can deal with these issues, such as MGA (Höhl et al. 2002), Multi-LAGAN (Brudno et al. 2003a), Shuffle-LAGAN (Brudno et al. 2003b), MultiPipMaker (Schwartz et al. 2003), MAVID (Bray and Pachter 2004), TBA (Blanchette et al. 2004), Mauve (Darling et al. 2004), MCAlign (Keightley and Johnson 2004), the CHAOS–DiAlign–ABC suite (Pöhler et al. 2005), MAP2 (Ye and Huang 2005) and AuberGene (Szklarczyk and Heringa 2006). There is clearly no dearth of algorithms available (and even more if you include pairwise-only methods), or of people willing to develop new ones. It would be interesting to compare all of these programs on a set of phylogenetic alignments (cf. Pollard et al. 2004 compared pairwise alignments only), but that is beyond the scope here. It is likely, however, that in their attempt to deal with whole genomes these algorithms sacrifice attention to details at the residue level, and for distantly related sequences this is precisely the level where the historical signal is located.

The basic objective of alignment, as conceived here, is to construct a detailed scenario that turns a single ancestral sequence into the set of contemporary sequences. This problem has long been considered for tandem repeats (see Bertrand and Gascuel 2005). Indeed, Sammeth and Stoye (2006) have developed a pairwise alignment model explicitly based on substitutions + indels + duplications + excisions. Thus, progress is being made algorithmically, although there is clearly a long way to go before the other processes can be included, and before we can proceed to multiple alignments.

The most obvious heuristic approach to phylogenetic alignment, based on the framework presented here, is to generate a list of possible biological events and then use this to put bounds on the alignment space. The steps might include:

Step (1) Identify the major sequence features, particularly in the context of the structure of the encoded product (if any). This can involve: (a) locating the structural elements in each sequence; (b) inferring the addition or absence of structural elements among the sequences, and any major changes in size; (c) locating conserved motifs, especially if they represent functional sites; (d) detecting repeats, whether tandem or otherwise, in single sequences; and (e) detecting inversions, translocations, etc., which require sequence comparison (e.g., pairwise alignments). Possible automated strategies are either discussed by Morrison (2006) or are discussed above. The main objective of this step is to accumulate the relevant biological data, so that the alignment is not simply a computerized string-matching exercise. It defines the sequence “fragments” that are possibly homologous, and for which the historical scenarios are then to be constructed. This explicitly puts biology into the alignment procedure by making the study of biological events the first step.

Step (2) Use similarity-based procedures to align well-conserved regions (i.e., maximize similarity). This can use local alignment procedures or motif-finding algorithms.

These regions then act as “anchors”, defining the boundaries between length–variable regions. Alternatively, closely related sequences can usually be aligned using similarity as the sole criterion.

Step (3) Use rule-based procedures to align the non-conserved regions or distantly related sequences, based on known or expected molecular processes (i.e., parsimony of events). These rules need to specify how to deal with changes in order or orientation of the sequence elements (e.g., transpositions and inversions), as well as with length changes (e.g., duplications, deletions, insertions). For example, aligning repeats can involve decisions about which copies to align against each other (Fig. 2), while aligning inversions involves decisions about whether to reverse complement some of the sequences or to stagger the alignment (Fig. 3). Both of these decisions will hopefully rely on the historical scenarios derived in step (1), but they may otherwise need an arbitrary convention.

This global alignment puts all of the information from step (1) together, using some weighting scheme to compare the different events, so that there is an optimality criterion (otherwise the output will be a trivial alignment with only fully conserved positions aligned). One contemporary example of this approach is the computer program T-Coffee (Notredame et al. 2000), which uses consistency as a weighting criterion to combine a set of pairwise alignments into a multiple alignment. We could thus construct a large library of pairwise alignments based on the evidence from step (1) and then use T-Coffee to combine the alignments. Unfortunately, the practical difficulty is that this program is very memory- and processor-intensive, which makes it impractical for most phylogenetic data sets.

Step (4) Check the alignment with respect to other criteria to ensure biological plausibility. This might include making sure that each individual event is aligned separately (i.e., no partial overlap), that there is consistency with respect to secondary structure, or that potential frame-shifts in protein-coding sequences have been identified. No fully automated help is currently available for this step, and so it must be based on manual checking.

One recent attempt to partly automate such a procedure for amino-acid sequences is the Promals program (Pei and Grishin 2007). This uses similarity to align very similar sequences into profiles, then uses third-order structure information to align the profiles, and then refines the alignment based on conserved blocks. In my experience this can work very well (after the nucleotides have been translated to amino acids), but the program often misses conserved motifs, for which the ProbCons program does much better; and so a combined approach is still needed for phylogenetic purposes.

If a particular alignment decision appears to be arbitrary (i.e., there is no evidence available from anywhere), such as it often is for variable numbers of microsatellite repeats, where the historical scenario may be ambiguous, then the sequences can be aligned using some convention. Possible conventions include (among others): (a) left-align variable numbers of repeats (e.g., slipped-strand mispairing suggests that the duplications are more likely to occur second in the sequence, so that blocks of residues can be aligned as far to the left as possible, with the gaps at the right); (b) prefer transitions to transversions (e.g., these are empirically observed to be more common); and (c) minimize the number of columns with mismatches (i.e., minimize the number of false positives) or minimize the number of mismatches per column (i.e., minimize the number of potentially informative mismatches). An example is shown in Fig. 2, where the original alignment of Kreitman (1983) right-aligns the repeated region whereas I have left-aligned it (as also in Fig. 4). Another convention is used in Fig. 6.

The use of conventions makes any procedure objective and repeatable, as it avoids variation among arbitrary choices., In practice, this is employed by all of the current computer algorithms, although this fact is not always made explicit (e.g., Clustal generally left-aligns blocks when there are several equally optimal choices, while MAFFT and POY right-align, and ProbCons centre-aligns). However, many researchers prefer to exclude regions where such conventions are necessary, on the grounds that the consequent evolutionary hypotheses are arbitrary (e.g., Kelchner 2000; Borsch et al. 2003, Löhne and Borsch 2005).

If we try to implement such an event-based alignment procedure, then we encounter the following practical problems: (1) how to score the alternative events, so that some mathematical function can be optimized (e.g., is one duplication worth more or less than one inversion?); (2) the number of alternative scenarios may mount up quickly. So, there are two basic problems. The first is biological: how to list all of the possible event types and their relative weights. The second is algorithmic: how to enumerate all of the possible optimal scenarios in practical computer time and space.

Issue (1) becomes particularly problematic when we consider possible combinations of events (e.g., is an inverted repeat worth the same as a duplication plus an inversion?) and sequence length (e.g., is two short inversions worth more or less than one long duplication?). Also, any weighting scheme may need to be both taxon-specific and gene-specific, in order to deal with observed differences in functional constraints. There is no biological basis for a uniform scheme, nor is there a mathematical basis for any other scheme; and the solution will be related to our conception of descriptive parsimony (see Ronquist 2003 for a discussion of weighting).

For issue (2), there are obvious algorithmic similarities to the problem of reconciling gene trees with a species tree (Charleston 1998) as well as to the alignment of RNA secondary structures. Experience with the latter two problems tells us that we may need to impose constraints on possible solutions in order to make the optimization feasible, as we will need to choose among many equally optimal solutions. These constraints can most obviously come from previous knowledge, such as the location of conserved motifs or gene boundaries (e.g., aligning genes separately in multi-gene analyses). It seems to me that far too much computer time is currently spent re-discovering alignment patterns identified in previous analyses of the same (or very similar) data sets, as well as assessing biologically unlikely alignments.

This leads to the issue of adding a new set of sequences to a pre-existing alignment, which Morrison (2006) refers to as jump-starting alignment and Kauff et al. (2007) as a ratchet. The objective here is to use the knowledge embodied in the previous alignment, so that effort is not repeated unnecessarily. Clearly, we need an effective method both to add the new sequences and to re-assess the old alignment, as the set of hypotheses contained in the original alignment may change with the addition of new sequences (i.e., the new sequences allow the previous evidence to be re-interpreted). The latter is an important issue that is not addressed when the new sequences are added based solely on similarity. One obvious example of implementing the above ideas, using protein-coding sequences, would be to have a computer program that will translate a DNA sequence into amino acids and then align it to an amino-acid profile, thus using second-order structure information (e.g., Goode and Rodrigo 2007). There are also editors that will use the secondary structure of each individual sequence to aid in the alignment of RNA- and non-coding sequences to a pre-existing alignment (e.g., Seibel et al. 2006). Furthermore, Kauff et al. (2007) discuss a general-purpose automated system for adding sequences to a profile, based on matching each new sequence to pre-defined conserved sequence blocks in the original ("core") alignment.

Quantifying alignment quality

A final important issue is the fact that we need some quantitative measurement of alignment quality (Vingron 1996), so that users can assess whether there are likely to have been problems (e.g., ambiguities) in constructing the alignment. This is particularly important given the range of objectives that exist for multiple sequence alignment (Talavera and Castresana 2007)—one could even define a "high-quality" alignment as being one that is simultaneously suitable for several objectives.

As a criterion we could, for example, specify the percent identity of a multiple alignment, given that alignment quality has a known relationship to this parameter (Morrison 2006). We would, however, need to standardize this measurement, as there are many ways to calculate it (May 2004). Alternatively, measures of reliability for multiple sequence alignments have been proposed based on a number of competing criteria (Pei and Grishin 2001; Thompson et al. 2001; Lassmann and Sonnhammer 2005; Ahola et al. 2006; Landan and Graur 2007; Ochoterena this volume). Unfortunately, there seems to be little consensus among these measurements as to what alignment "reliability" might mean.

Furthermore, it would be good to have some quantitative measurement of alignment quality at each position in the alignment, so that we can assess where any problems might have occurred. For example, probabilistic alignment methods have an inherent ability to provide a measure of reliability for each position in their own alignment, although this measure cannot necessarily be applied to independently derived alignments. This facility is provided by a number of computer programs (e.g., ProAlign, Löytynoja and Milinkovitch 2003; ProbCons, Do et al. 2005; Prank, Löytynoja and Goldman 2005; BAli-Phy, Suchard and Redelings 2006).

A number of more-general algorithms have been proposed for this purpose, but there apparently has been no comparative study of them. So, as a preliminary assessment I have directly compared some of them for a specific amino-acid data set in Fig. 10. There are basically two types of measurement: (1) quantitative scoring schemes, which provide a reliability score for each aligned position (Dopazo 1997; Thompson et al. 1997; Notredame et al. 1998; O'Brien and Higgins 1998; Pei and Grishin 2001), and (2) selection schemes, which select a subset of the aligned positions as being reliably aligned (Martin et al. 1995; Grundy and Naylor 1999; Castresana 2000; Löytynoja and Milinkovitch 2001; Thompson et al. 2001: Shan et al. 2003; Lawrence et al. 2004).

The multiple sequence alignment used in the example is quite a challenging one, as these FN1 amino-acid sequences have 34.2% average pairwise identity, with only four perfectly conserved alignment positions (all C, which form disulphide bridges), one almost conserved (Y), one conserved plus a gap (G), and three well-conserved (W, G, G). This situation is reflected in fairly small reliability scores for most of the positions from all four computer programs (Fig. 10). However, only WET and AL2CO have highly correlated scores, with a Spearman rank correlation of 0.724 (the other pairwise correlations are 0.372–0.595). There thus seems to be little agreement as to what a reliable alignment position might look like.

```
Gblocks (100% gaps)       111111111111111111111111111111111111111111
Gblocks  (50% gaps)       1111111111111                1111111111111
Gblocks   (0% gaps)
SOAP  (30% support)   111 111   1111111111111          111       111 111
SOAP  (90% support)   111 111           11111          111       111 111
NorMD (score > 0.3)     1 111    11111    11111      1111111     111 1
NorMD (score > 0.5)     1 111                11        111 1     111 1
OSA  (90% distance)        111111111111111111111111111
OSA  (80% distance)   111 11111111111111111111111111111
LumberJack            111 11111111111111111111111111

WET                   202 922200011802242091412003211229091403022209 002
AL2CO                 010 915014301814504271710125011109292716395219 110
T-Coffee              432 4334122345565-6565543232232234455443354334 345
ClustalX              221 903010111922203341611103112109291514110039 221

Consensus structure   --- -SSTTTTSS--TT--SSSSS-TTTT-SSSSS--TTTT---SS ---

FA12 HUMAN/135-170    KEK|CFEPQLLRFFHKN-EIWYRTEQ-AAVAR-CQCKG-PD--AHC|QRL
HGFA HUMAN/202-237    TEK|CFD-ETRYEYLEGGDRWARVRQ-GHVEQ-CECFG-GR--TWC|EGT
 TPA HUMAN/41-78      QVI|CRDEKTQMIYQQH-QSWLRPVLRSNRVEYCWC-NSGR--AQC|HSV
 TPA MOUSE/38-75      RAT|CRDEPTQTTYQQH-QSWLRPMLRSSRVEYCRC-NSGL--VQC|HSV
 TPA RAT/38-75        RAT|CRDEQTQTTYQQH-QSWLRPMLRGNRVEYCRC-NSGL--AQC|HSV
URT2 DESRO/42-79      GVA|CRDEKTQMIYQQQ-ESWLRPEVRSKRVEHCRC-DRGL--AQC|HTV
FINC BOVIN/21-56      KPG|CYD-N-GKHYQIN-QQWERTYL-GSALV-CTCYGGSRG-FNC|ESK
FINC HUMAN/52-87      KPG|CYD-N-GKHYQIN-QQWERTYL-GNALV-CTCYGGSRG-FNC|ESK
FINC BOVIN/66-104     EET|CFDKYTGNTYRVG-DTYERPKD-SMIWD-CTCIGAGRGRISC|TIA
FINC HUMAN/97-135     EET|CFDKYTGNTYRVG-DTYERPKD-SMIWD-CTCIGAGRGRISC|TIA
FINC BOVIN/110-148    ANR|CHE-G-GQSYKIG-DTWRRPHETGGYMLECVCLGNGKGEWTC|KPI
FINC HUMAN/141-179    ANR|CHE-G-GQSYKIG-DTWRRPHETGGYMLECVCLGNGKGEWTC|KPI
FINC BOVIN/155-194    AEK|CFDQAAGTSYVVG-ETWEKPYQ-GWMMVDCTCLGEGSGRITC|TSR
FINC HUMAN/186-225    AEK|CFDHAAGTSYVVG-ETWEKPYQ-GWMMVDCTCLGEGSGRITC|TSR
FINC BOVIN/200-239    RNR|CNDQDTRTSYRIG-DTWSKKDNRGNLLQ-CICTGNGRGEWKC|ERH
FINC HUMAN/231-270    RNR|CNDQDTRTSYRIG-DTWSKKDNRGNLLQ-CICTGNGRGEWKC|ERH
FINC BOVIN/277-311    YGH|CVT-DSGVVYSVG-MQWLKTQG-NKQML-CTCLGNG---VSC|QET
FINC HUMAN/308-342    YGH|CVT-DSGVVYSVG-MQWLKTQG-NKQML-CTCLGNG---VSC|QET
FINC BOVIN/439-477    EEI|CTT-NEGVMYRIG-DQWDKQHDMGHMMR-CTCVGNGRGEWTC|VAY
FINC HUMAN/470-508    EEI|CTT-NEGVMYRIG-DQWDKQHDMGHMMR-CTCVGNGRGEWTC|IAY
FINC BOVIN/487-524    RDQ|CIV-D-GITYNVN-DTFHKRHEEGHMLN-CTCFGQGRGRWKC|DPV
FINC HUMAN/518-555    RDQ|CIV-D-DITYNVN-DTFHKRHEEGHMLN-CTCFGQGRGRWKC|DPV
FINC BOVIN/530-568    VDQ|CQDSETRTFYQIG-DSWEKYLQ-GVRYQ-CYCYGRGIGEWAC|QPL
FINC HUMAN/561-599    VDQ|CQDSETGTFYQIG-DSWEKYVH-GVRYQ-CYCYGRGIGEWHC|QPL
FINC BOVIN/2085-2124  DDS|CFDPYTVSHYAIG-EEWERLSDSGFKLS-CQCLGFGSGHFRC|DSS
FINC HUMAN/2206-2245  DDS|CFDPYTVSHYAVG-DEWERMSESGFKLL-CQCLGFGSGHFRC|DSS
FINC BOVIN/2130-2167  SKW|CHD-N-GVNYKIG-EKWDRQGENGQMMS-CTCLGNGKGEFKC|DPH
FINC HUMAN/2251-2288  SRW|CHD-N-GVNYKIG-EKWDRQGENGQMMS-CTCLGNGKGEFKC|DPH
FINC BOVIN/2174-2209  EAT|CYD-D-GKTYHVG-EQWQKEYL-GAICS-CTCFGGQRG-WRC|DNC
FINC HUMAN/2295-2330  EAT|CYD-D-GKTYHVG-EQWQKEYL-GAICS-CTCFGGQRG-WRC|DNC
```

Fig. 10 Aligned domain and flanking regions of the fibronectin type I repeat (*FN1*) of 30 amino-acid sequences from several mammalian genes. The *first ten lines* show the blocks selected by five different methods as suitable for phylogenetic analysis. The *next four lines* show positional conservation as determined by four different scoring schemes. The *next line* shows the consensus secondary structure of the protein domain (see Baron et al. 1990): *S* H-bonded beta-strand or isolated beta-bridge, *T* bend or H-bonded turn, – random coil. The *vertical bars* delimit the FN1 domain. The alignment is modified from that in the Pfam (Finn et al. 2006) database (Version 17.0, domain family PF00039): *FA12* coagulation factor XII, *FINC* fibronectin, *HGFA* hepatocyte growth factor activator, *TPA* tissue-type plasminogen activator, *URT2* salivary plasminogen activator. The Gblocks v.0.91b (Castresana 2000) "selected blocks" are based on conservation of identity; the default settings were used, with three different alternatives assessed for the percentage of sequences containing gaps (note that no positions were selected when no sequences were allowed to have gaps). The SOAP v.1.2a4 (Löytynoja and Milinkovitch 2001) "stability" is measured with respect to variation in the Clustal gap-opening and gap-extension penalties; the settings used were GOP = 2.5,5,10,20,40, GEP = 0.05,0.1,0.2,0.4,0.8, with two alternative cutoffs of support. The NorMD v.1.0 (Thompson et al. 2001) "normalized mean distance" is based on pairwise distances; the settings used were GOP = 1, GEP = 0.1, window = 5, Gonnet250 matrix, with two alternative score cut-offs; the overall score of the alignment is 0.507. OSA v.2.0 (Martin et al. 1995) locates "smallest blocks" with similar pairwise genetic distances to the whole alignment; the settings used were 5,000 bootstrap replicates, Jukes-Cantor distance, gaps excluded pairwise, no length correction, and two different distances assessed as cut-offs. LumberJack v.7.10 (Lawrence et al. 2004) identifies blocks that have their phylogenetic tree being most similar to that of the whole alignment; the default settings were used. The WET v.1.3 (Dopazo 1997) "evolutionary index" is based on conservativeness of amino acid differences as predicted from nucleotide differences; italic values represent negative variability scores. The AL2CO v.1.2 (Pei and Grishin 2001) "conservation" is based on weighted entropy. The T-Coffee v.3.27 (Notredame et al. 2000) score represents "consistency" among global and local alignments. The ClustalX v1.83 (Thompson et al. 1997) "quality" is based on conservativeness of amino acid differences; 0 = 0–9, 1 = 10–19,…, 9 = 90–100

As far as the selection procedures are concerned, LumberJack and OSA find only the first two-thirds of the alignment to be reliable, NorMD and SOAP find several scattered reliable regions, and Gblocks restricts its choice to the domain itself (Fig. 10). There is thus little consensus here, either, with three distinct concepts as to which positions are worthy of further analysis.

In this context, it might be helpful to have some method of visually comparing several multiple alignments, either as a replacement for, or an adjunct to, the quantitative scores. A step in this direction has been taken by Shih et al. (2006), although their program is designed for a few long sequences rather than many sequences. It is, of course, the rapidly growing number of sequences that it is of increasing concern for phylogenetic studies. However, it is not clear that phylogenetic analyses will need to extend beyond a couple of hundred sequences, as it is probably a more efficient strategy to use exemplars to assess the main cladistic structure within a clade and then to examine each subclade in detail separately.

References

Ahola V, Aittokallio T, Vihinen M, Uusipaikka E (2006) A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinform 7:484

Baron M, Norman D, Willis A, Campbell ID (1990) Structure of the fibronectin type I module. Nature 345:642–646

Barta JR (1997) Investigating phylogenetic relationships within the Apicomplexa using sequence data: the search for homology. Methods 13:81–88

Beebe NW, Cooper RD, Morrison DA, Ellis JT (2000) Subset partitioning of the ribosomal DNA small subunit and its effects on the phylogeny of the *Anopheles punctulatus* group. Insect Molec Biol 9:515–520

Bertrand D, Gascuel O (2005) Topological rearrangements and local search method for tandem duplication trees. IEEE/ACM Trans Comput Biol Bioinform 2:15–28

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the Threaded Blockset Aligner. Genome Res 14:708–715

Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W (2003) Noncoding plastid *trnT–trnF* sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol 16:558–576

Borsch T, Hilu KW, Wiersema JH, Löhne C, Barthlott W, Wilde V (2007) Phylogeny of *Nymphaea* (Nymphaeaceae): evidence from substitutions and microstructural changes in the chloroplast *trnT–trnF* region. Int J Plant Sci 168:639–671

Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699

Brower AVZ, Schawaroch V (1996) Three steps of homology assessment. Cladistics 12:265–272

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S (2003a) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S (2003b) Glocal alignment: finding rearrangements during alignment. Bioinformatics 19:i54–i62

Cammarano P, Creti R, Sanangelantoni AM, Palm P (1999) The Archaea monophyly issue: a phylogeny of translational elongation factor g(2) sequences inferred from an optimized selection of alignment positions. J Molec Evol 49:524–537

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang Z, Yu N, Gutell RR (2002) The comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinform 3:2

Cartmill M (1994) A critique of homology as a morphological concept. Am J Physical Anthropol 94:115–123

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molec Biol Evol 17:540–552

Charleston MA (1998) Jungles: a new solution to the host/parasite phylogeny reconciliation problem. Math Biosci 149:191–223

Colbourn CJ, Kumar S (2007) Lower bounds on multiple sequence alignment using exact 3-way alignment. BMC Bioinform 8:140

Creer S (2007) Choosing and using introns in molecular phylogenetics. Evol Bioinform 3:99–108

Damberger SH, Gutell RR (1994) A comparative database of group I intron structures. Nucleic Acids Res 22:3508–3510

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403

Dessimoz C, Cannarozzi GM, Gil M, Margadant D, Roth A, Schneider A, Gonnet GH (2005) OMA, A comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. Lect Notes Comput Sci 3678:61–72

de Pinna MCC (1991) Concepts and tests of homology in the cladistic paradigm. Cladistics 7:367–394

Dewey CN, Pachter L (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. Human Molec Genet 15:R51–R56

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340

Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teacher 35:125–129

Dopazo J (1997) A new index to find regions showing an unexpected variability or conservation in sequence alignments. Comput Appl Biosc 13:313–317

Du Z, Lin F (2007) Pattern-constrained multiple polypeptide sequence alignment. Comput Biol Chem 29:303–307

Ellis J, Morrison D (1995) Effects of sequence alignment on the phylogeny of *Sarcocystis* deduced from 18S rDNA sequences. Parasitol Res 81:696–699

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251

Fleissner R, Metzler D, von Haeseler A (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst Biol 54:548–561

Frith MC, Hansen U, Spouge JL, Weng Z (2004) Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 32:189–200

Gillespie JJ (2004) Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules. Molec Phylogenet Evol 33:936–943

Gillespie JJ, Yoder MJ, Wharton RA (2005) Predicted secondary structure for 28S and 18S rRNA from Ichneumonoidea (Insecta:Hymenoptera:Apocrita): impact on sequence alignment and phylogeny estimation. J Molec Evol 61:114–137

Giribet G, Edgecombe GD, Wheeler WC (2001) Arthropod phylogeny based on eight molecular loci and morphology. Nature 413:157–160

Golenberg EM, Clegg MT, Durbin ML, Doebley J, Ma DP (1993) Evolution of a noncoding region of the chloroplast genome. Molec Phylogen Evol 2:52–64

Golubchik T, Wise MJ, Eastel S, Jermiin LS (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol Biol Evol 24:2433–2442

Goode MG, Rodrigo AG (2007) SQUINT: a multiple alignment program and editor. Bioinformatics 23:1553–1555

Graham SW, Reeves PA, Burns ACE, Olmstead RG (2000) Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. Int J Plant Sci 161:S83–S96

Grundy WN, Naylor GJP (1999) Phylogenetic inference from conserved alignments. J Exp Zool 285:128–139

He Y, Jones J, Armstrong M, Lamberti F, Moens M (2005) The mitochondrial genome of *Xiphinema americanum sensu stricto* (Nematoda: Enoplea): considerable economization in the length and structural features of encoded genes. J Molec Evol 61:819–833

Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 163:GC17–GC26

Hertwig S, de Sá RO, Haas A (2004) Phylogenetic signal and the utility of 12S and 16S mtDNA in frog phylogeny. J Zool Syst Evol Res 42:2–18

Hickson RE, Simon C, Cooper A, Spicer GS, Sullivan J, Penny D (1996) Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. Molec Biol Evol 13:150–169

Höhl M, Kurtz S, Ohlebusch E (2002) Efficient multiple genome alignment. Bioinformatics 18:S312–S320

Höhl M, Ragan MA (2007) Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol 56:206–221

Hoot SB, Douglas AW (1998) Phylogeny of the Proteaceae based on *atp*B and *atp*B–*rbcL* intergenic spacer region sequences. Aust Syst Bot 11:301–320

Jansen RK, Kaittanis C, Saski C, Lee S-B, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analysis of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol. Biol 6:32

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol 53:638–643

Johnson R (1982) Parsimony principles in phylogenetic systematics: a critical re-appraisal. Evol Theory 6:79–90

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518

Kauff F, Cox CJ, Lutzoni F (2007) WASABI: an automated sequence processing system for multigene phylogenies. Syst Biol 56:523–531

Keightley PD, Johnson T (2004) MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. Genome Res 14:442–450

Kelchner SA (2000) The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann Missouri Bot Gard 87:482–498

Kelchner SA (2002) Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. Amer J Bot 89:1651–1669

Kelchner SA, Clark LG (1997) Molecular evolution and phylogenetic utility of the chloroplast *rpl*16 intron in *Chusquea* and the Bambusoideae (Poaceae). Molec Phylogenet Evol 8:385–397

Kelchner SA, Wendel JF (1996) Hairpins create minute inversions in non-coding regions of chloroplast DNA. Curr Genet 30:259–262

Kellogg EA, Juliano ND (1997) The structure and function of RuBisCo and their implications for systematic studies. Am J Bot 84:413–428

Kim J, Sinha S (2007) Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. Bioinformatics 23:289–297

Kiryu H, Tabei Y, Kin T, Asai K (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. Bioinformatics 23:1588–1598

Kjer KM (1995) Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Molec Phylogenet Evol 4:314–330

Kjer KM (1997) An alignment template for amphibian 12S rRNA, domain III: conserved primary and secondary structural motifs. J. Herpetol 31:599–604

Kjer KM, Baldridge GD, Fallon AM (1994) Mosquito large subunit ribosomal RNA: simultaneous alignment of primary and secondary structure. Biochim Biophys Acta 1217:147–155

Kjer KM, Gillespie JJ, Ober KA (2006) Structural homology in ribosomal RNA, and a deliberation on POY. Arthropod Syst Phylogeny 64:159–164

Kjer KM, Gillespie JJ, Ober KA (2007) Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. Syst Biol 56:133–146

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

Kumar S, Filipski A (2007) Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res 17:127–135

Lambert C, Van Campenhout J-M, DeBolle X, Depiereux E (2003) Review of common sequence alignment methods: clues to enhance reliability. Curr Genom 4:131–146

Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments. Molec Biol Evol 24:1380–1383

Lassmann T, Sonnhammer ELL (2005) Automatic assessment of alignment quality. Nucleic Acids Res 33:7120–7128

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208–214

Lawrence CJ, Zmasek CM, Dawe RK, Malmberg RL (2004) LumberJack: a heuristic tool for sequence alignment exploration and phylogenetic inference. Bioinformatics 20:1977–1979

Lebrun E, Santini JM, Brugna M, Ducluzeau A-L, Ouchane S, Schoepp-Cothenet B, Baymann F, Nitschke W (2006) The rieske protein: a case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction. Molec Biol Evol 23:1180–1191

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork O (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res 32:142–144

Ljunggren EL, Bergström K, Morrison DA, Mattsson JG (2006) Characterisation of an atypical antigen from *Sarcoptes scabiei* containing an MADF domain. Parasitology 132:117–126

Löhne C, Borsch T (2005) Molecular evolution and phylogenetic utility of the *petD* Group II intron: a case study in basal angiosperms. Molec Biol Evol 22:317–332

Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA 102:10557–10562

Löytynoja A, Milinkovitch MC (2001) SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics 17:573–574

Löytynoja A, Milinkovitch MC (2003) A hidden markov model for progressive multiple alignment. Bioinformatics 19:1505–1513

Lunter G (2007) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. Bioinformatics 23:i289–i296

Lunter G, Drummond AJ, Miklós I, Hein J (2005) Statistical alignment: recent progress, new applications, and challenges.

In: Nielsen R (ed) Statistical methods in molecular evolution. Springer, New York, pp 375–405

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki C, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005) CDD: a conserved domain database for protein classification. Nucleic Acids Res 33:D192–D196

Martin MJ, González-Candelas F, Sobrino F, Dopazo J (1995) A method for determining the position and size of optimal sequence regions for phylogenetic analysis. J Molec Evol 41:1128–1138

Martin W, Roettger M, Lockhart PJ (2007) A reality check for alignments and trees. Trends Genet 23:478–480

May ACW (2004) Percent sequence identity: the need to be explicit. Structure 12:737–738

Messer PW, Arndt PF (2007) The majority of recent short DNA insertions in the human genome are tandem duplications. Molec Biol Evol 24:1190–1197

Mishler BD (2005) The logic of the data matrix in phylogenetic analysis. In: Albert VA (ed) Parsimony, phylogeny, and genomics. Oxford University Press, Oxford, pp 57–70

Morrison DA (2006) Multiple sequence alignment for phylogenetic purposes. Aust Syst Bot 19:479–539

Morrison DA, Bornstein S, Thebo P, Wernery U, Kinne J, Mattsson JG (2004) The current status of the small subunit rRNA: phylogeny of the coccidia (Sporozoa). Int J Parasitol 34:501–514

Morrison DA, Ellis JT (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. Molec Biol Evol 14:428–441

Müller K, Borsch T (2005a) Phylogenetics of *Utricularia* (Lentibulariaceae) and molecular evolution of the *trn*K intron in a lineage with high substitutional rates. Plant Syst Evol 250:39–67

Müller K, Borsch T (2005b) Phylogenetics of Amaranthaceae based on *mat*K/*trn*K sequence data—evidence from parsimony, likelihood, and Bayesian methods. Ann Missouri Bot Gard 92:66–102

Mugridge NB, Morrison DA, Jäkel T, Heckeroth AR, Tenter AM, Johnson AM (2000) Effects of sequence alignment and structural domains of ribosomal DNA on phylogeny reconstruction for the protozoan family Sarcocystidae. Molec Biol Evol 17:1842–1853

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Molec Biol 302:205–217

Notredame C, Holm L, Higgins DG (1998) COFFEE: an objective function for multiple sequence alignments. Bioinformatics 14:407–422

O'Brien EA, Higgins DG (1998) Empirical estimation of the reliability of ribosomal RNA alignments. Bioinformatics 14:830–838

O'Dushlaine CT, Shields DC (2006) Tools for the identification of variable and potentially variable tandem repeats. BMC Genom 7:290

Ogden TH, Rosenberg MS (2006) Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol 55:314–328

Ogden TH, Rosenberg MS (2007) Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. Syst Biol 56:182–193

Papadopoulos JS, Agarwala R (2007) COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics 23:1073–1079

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

Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712

Pei J, Grishin NV (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23:802–808

Phillips A, Janies D, Wheeler W (2000) Multiple sequence alignment in phylogenetic analysis. Molec Phylogen Evol 16:317–330

Phuong TM, Do CB, Edgar RC, Batzoglou S (2006) Multiple alignment of protein sequences with repeats and rearrangements. Nucleic Acids Res 34:5932–5942

Pöhler D, Werner N, Steinkamp R, Morgenstern B (2005) Multiple alignment of genomic sequences using CHAOS, DIALIGN and ABC. Nucleic Acids Res 33:W532–W534

Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB (2004) Benchmarking tools for the alignment of functional noncoding DNA. BMC Bioinform 5:6

Pons J, Vogler AP (2006) Size, frequency, and phylogenetic signal of multiple-residue indels in sequence alignment of introns. Cladistics 22:144–156

Prychitko TM, Moore WS (2003) Alignment and phylogenetic analysis of β-fibrinogen intron 7 sequences among avian orders reveal conserved regions within the intron. Mol Biol Evol 20:762–771

Quandt D, Müller K, Huttunen S (2003) Characterisation of the chloroplast DNA *psb*T-H region and the influence of dyad symmetrical events on phylogenetic reconstructions. Pl Biol 5:400–410

Quandt D, Müller K, Stech M, Frahm J-P, Frey W, Hiku KW, Borsch T (2004) Molecular evolution of the chloroplast *trn*L-F region in land plants. In: Goffinet B, Hollowell V, Magill R (eds) Molecular systematics of bryophytes. Missouri Botanical Garden Press, St Louis, pp 13–37

Raphael B, Zhi D, Tang H, Pevzner P (2004) A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res 14:2336–2346

Redelings BD, Suchard MA (2005) Joint bayesian estimation of alignment and phylogeny. Syst Biol 54:401–418

Ronquist F (2003) Parsimony analysis of coevolving species associations. In: Page RDM (ed) Phylogeny, cospeciation and evolution. University of Chicago Press, Chicago, pp 22–64

Sammeth M, Heringa J (2006) Global multiple-sequence alignment with repeats. Proteins Struct Funct Bioinform 64:263–274

Sammeth M, Stoye J (2006) Comparing tandem repeats with duplications and excisions of variable degree. IEEE/ACM Trans Computat Biol Bioinform 3:395–407

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

Sanson GFO, Kawashita SY, Brunstein A, Briones MRS (2002) Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. Mol Biol Evol 19:170–178

Schultz J, Maisel S, Gerlach D, Müller T, Wolf M (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA 11:361–364

Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A (2003) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res 31:3518–3524

Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M (2006) 4SALE—a tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinform 7:498

Shan Y, Milios EE, Roger AJ, Blouin C, Susko E (2003) Automatic recognition of regions of intrinsically poor multiple alignment using machine learning. In: Proceedings of the IEEE computer society second conference in bioinformatics (CSB'03). IEEE Press, Piscataway, pp 482–483

Shih AC-C, Lee DT, Lin L, Peng C-L, Chen S-H, Wu Y-W, Wong C-Y, Chou M-Y, Shiao T-C, Hsieh M-F (2006) SinicView: a visualization environment for comparisons of multiple nucleotide sequence alignment tools. BMC Bioinform 7:103

Simmons MP (2004) Independence of alignment and tree search. Molec Phylogenet Evol 31:874–879

Smith HO, Annau TM, Chandrasegaran S (1990) Finding sequence motifs in groups of functionally related proteins. Proc Natl Acad Sci USA 87:826–830

Stebbings LA, Mizuguchi K (2004) HOMSTRAD: recent developments of the homologous protein structure alignment database. Nucleic Acids Res 32:D203–D207

Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous bayesian inference of alignment and phylogeny. Bioinformatics 22:2047–2048

Szklarczyk R, Heringa J (2006) AuberGene—a sensitive genome alignment tool. Bioinformatics 22:1431–1436

Szymanski M, Erdmann VA, Barciszewski J (2007) Noncoding RNAs database (ncRNAdb). Nucleic Acids Res 35:D162–D164

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Thompson JD, Plewniak F, Ripp R, Thierry J-C, Poch O (2001) Towards a reliable objective function for multiple sequence alignments. J Molec Biol 314:937–951

Torarinsson E, Havgaard JH, Gorodkin J (2007) Multiple structural alignment and clustering of RNA sequences. Bioinformatics 23:926–932

Vingron M (1996) Near-optimal sequence alignment. Curr Opin Struct Biol 6:346–352

Wegner K, Jansen S, Wuchty S, Gauges R, Kummer U (2004) CombAlign: a protein sequence comparison algorithm considering recombinations. In Silico Biol 4:0021

Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Res 34:D327–D331

Wheeler TJ, Kececioglu JD (2007) Multiple alignment by aligning alignments. Bioinformatics 23:i559–i568

Wheeler W (1996) Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12:1–9

Wheeler WC (1999) Fixed character states and the optimization of molecular sequence data. Cladistics 15:379–385

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

Wilm A, Mainz I, Steger G (2006) An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Molec Biol 1:19

Xiao L, Sulaiman IM, Ryan UM, Zhou L, Atwill ER, Tischler ML, Zhang X, Fayer R, Lal AA (2002) Host adaptation and host-parasite co-evolution in *Cryptosporidium*: implications for taxonomy and public health. Int J Parasitol 32:1773–1785

Xu X, Ji Y, Stormo GD (2007) RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. Bioinformatics 23:1883–1891

Yao Z, Weinberg Z, Ruzzo WL (2006) CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 22:445–452

Ye L, Huang X (2005) MAP2: multiple alignment of syntenic genomic sequences. Nucleic Acids Res 33:162–170