CSIRO PUBLISHING
www.publish.csiro.au/journals/asb *Australian Systematic Botany* **19**, 479–539

L. A. S. JOHNSON REVIEW No. 8

Multiple sequence alignment for phylogenetic purposes

David A. Morrison

Department of Parasitology (SWEPAR), National Veterinary Institute and Swedish University of Agricultural Sciences, 751 89 Uppsala, Sweden. Email: David.Morrison@bvf.slu.se

Contents

Abstract. I have addressed the biological rather than bioinformatics aspects of molecular sequence alignment by covering a series of topics that have been under-valued, particularly within the context of phylogenetic analysis. First, phylogenetic analysis is only one of the many objectives of sequence alignment, and the most appropriate multiple alignment may not be the same for all of these purposes. Phylogenetic alignment thus occupies a specific place within a broader context. Second, homology assessment plays an intricate role in phylogenetic analysis, with sequence alignment consisting of primary homology assessment and tree building being secondary homology assessment. The objective of phylogenetic alignment thus distinguishes it from other sorts of alignment. Third, I summarise what is known about the serious limitations of using phenetic similarity as a criterion for automated multiple alignment, and provide an overview of what is currently being done to improve these computerised procedures. This synthesises information that is apparently not widely known among phylogeneticists. Fourth, I then consider the recent development of automated procedures for combining alignment and tree building, thus integrating primary and secondary homology assessment. Finally, I outline various strategies for increasing the biological content of sequence alignment procedures, which consists of taking into account known evolutionary processes when making alignment decisions. These procedures can be objective and repeatable, and can involve computerised algorithms to automate much of the work. Perhaps the most important suggestion is that alignment should be seen as a process where new sequences are added to a pre-existing alignment that has been manually curated by the biologist.

Introduction

Sequence alignment is a fascinating subject. Unfortunately, what is often most fascinating about it is not the actual topic itself but, rather, the somewhat cavalier way that so many molecular biologists treat it. They will spend a huge amount of time collecting their data, and then potentially throw away

© CSIRO 14 December 2006 10.1071/SB06020 1030-1887/06/060479

all of their good work by feeding the data into a computer program with default parameter settings, apparently trusting the outcome to good luck rather than to good management. This attitude has always seemed incomprehensible to me (and therefore perversely fascinating) because I have always believed that there should be as much biology in a sequence alignment as there is mathematics and computing. This review is about just that idea, and is dedicated to putting biology back into the sequence alignment procedures used as part of phylogenetic analyses. The primary importance of biological insight in phylogenetic studies is something that L. A. S. Johnson was very insistent upon; and this topic is thus one of which I feel he would have approved.

Alignment is all about mapping the relationships between residues in a set of molecular sequences. These sequences could be nucleotide (DNA) sequences, or they could be products of the DNA such as RNA or amino acid sequences. They could be coding gene sequences such as proteins or rRNA, or they could be non-coding sequences such as introns or spacers. Whatever form the sequences take, before we can further analyse them we need to map the one-to-one relationships between the residues, so that we are comparing like with like. If the sequences are almost identical then this process may be unproblematic, especially if the sequences are almost identical in length. However, as the percentage identity decreases the process of alignment becomes more problematic, particularly when the lengths are unequal, as this means that gaps must be introduced into one or more of the sequences in order to equalise the lengths. (Note that gaps may also need to be added to equal-length sequences, if this helps map the relationships). Some form of quantitative procedure is needed, in order to produce reliable and repeatable alignments.

Relative to (say) studies of sequence alignment for protein structure prediction or database searching, much less attention seems to have been paid to date to problems of sequence alignment in phylogenetic studies, except to note that the problems of insertions and deletions (indels) can make alignment of sequences (or parts of sequences) not only difficult, but sometimes impossible. However, these alignment issues are at least as important as are other problems in phylogenetic tree building (e.g. sequence length, tree-inference methods, compositional bias, site-to-site variation) because they are fundamental to the concepts of character and character-state homology. More to the point, establishing homology for molecular characters may actually be harder than for other types of characters.

The alignment process is thus seen by many biologists as being a bioinformatics issue rather than a biological one. Indeed, sequence alignment is often claimed to be one of the major 'open' problems in computational biology (Karp 2002; Greenberg *et al*. 2004). This is why the alignment process is usually left to a computer program, perhaps with some *post hoc* re-alignment by eye. However, no-one has yet succeeded in putting much biological insight into sequence alignment programs, although this is not for want of trying. Therefore this insight must come directly from the biologist, who needs to pay careful attention to the rationale for the alignment decisions—the mathematics and computing are there to help, but not to replace, the biology.

Mathematics is about manipulating symbols, without reference to the objects being symbolised; that is its universal strength. So, algorithmically, sequence alignment is about lining up 'strings' of 'letters' into 'columns' by padding them out with 'gaps', and no heed needs to be paid to what the strings, letters, columns or gaps represent. However, biologically, sequence alignment is about establishing relationships between physical characteristics of real organisms; and to a biologist the characteristics of the objects being symbolised take precedence over the symbols themselves. In phylogeny, we do not align letters into columns but instead establish homology between characters and their states. This is not a trivial semantic issue, because many of the algorithms used for sequence alignment have precious little relevance to establishing homology, even if they do line up the letters quite neatly.

Putting biology into an alignment consists of thinking about the biological processes that you are postulating must have occurred in order to generate the alignment in the first place, and making these hypotheses plausible as well as parsimonious. Many alignments may look superficially very similar, and indeed they may behave identically in any specified data analysis, but that does not mean that they are equally plausible biologically. Alignments are representations of evolutionary history, and as biologists we cannot accept implausible hypotheses of evolutionary events, even if they seem not to affect the immediate data analyses. For example, a choice between two similar alignments may not affect construction of a phylogenetic tree but they can have consequences when we come to consider character evolution on that tree (e.g. the origin of molecular functions).

Our alignment procedures are thus an imperfect attempt to reconstruct unknowable evolutionary events, and to represent those events in a particular format. In many ways this is the hardest thing that a biologist can try to do, because the patterns being examined are unobservable historical ones rather than contemporary empirical observations. Consequently, the only protection that we can have against false conclusions is the quality of the data analysis. An alignment is only as good as the steps taken to ensure the highest quality of data and to evaluate and use the most appropriate method for the data analysis. Most scientists are very aware of (and try to avoid) the 'garbage in, garbage out' phenomenon, but it is equally possible to put useful things in and still get garbage out, if the processing is inappropriate. This will happen if biology ceases to be a science and becomes instead a series of algorithms. Our objective should be biological plausibility rather than mathematical optimality.

Since there is no simple way to turn our biological criterion into a mathematical one (i.e. it is difficult to give a formal definition of the biological task of studying evolutionary history that could then be turned into an optimisation problem with a mathematical solution; Vingron 1999), we cannot unthinkingly trust our alignment decisions solely to a computer algorithm.

I start this review of these topics by pointing out that the use of multiple alignments for phylogenetic purposes is only one of the many objectives of sequence alignment, and that the most appropriate alignment may not be the same for all of these purposes; this places phylogenetic alignment within a broader context. I then proceed to emphasise the intricate role that homology assessment plays in phylogenetic analyses, with sequence alignment consisting of primary homology assessment and tree building being secondary homology assessment; this provides the theoretical background for understanding the objective of phylogenetic alignment. I then summarise what is known about the serious limitations of using phenetic similarity as a criterion for producing automated multiple alignments, before proceeding to an overview of what is currently being done to improve these computerised procedures; this synthesises information that is apparently not widely known among phylogeneticists. Then follows some consideration of the recent development of automated procedures for combining alignment and tree building, thus integrating primary and secondary homology assessment. I then finish by outlining various strategies for increasing the biological insight that is used for sequence alignment procedures; this is the most important part of the review.

The 4 × 4 = 16 (and more) types of sequence alignment

There is a strong tendency for biologists to speak of 'sequence alignment' as though it is a single concept. However, there are four distinctly different objectives for sequence alignment and there are four different modes of sequence alignment, making a total of 16 different types, none of which necessarily entails the same alignment for any particular set of sequences. Moreover, the most commonly used type of alignment in practice is referred to as an 'optimal alignment', which may refer to optimisation of either global or local similarity; and this will probably be a suboptimal progressive alignment in practice anyway. So, sequence alignment is even more complex than my simple 16-group classification. It is my intention in this introductory section to highlight the difference between these various types of alignment, and to illustrate why they may actually all be mutually exclusive in practice. This will put alignment for phylogenetic purposes into context, and emphasise that it is a specialist procedure within a much broader field.

My illustrative example concerns a small section of the amino-acid sequence of the metallo-β-lactamase protein domain-superfamily (also known as Lactamase B in many of the protein databases), for a set of six species for which the protein structures are known (Fig. 1). The average pairwise amino-acid identity among the six full sequences is 21%, which is within what is known as the 'twilight zone' of protein similarity, where sequence alignment can be very problematic. The two most different sequences, human hydroxyacylglutathione hydrolase (or glyoxalase II) (labelled 1qh5A in the figure) and a bacterial penicillinase (labelled 1smlA), have only 16% amino-acid identity. Needless to say, I have deliberately chosen this example because it neatly illustrates the points that I wish to make—I do not suggest that anyone would necessarily wish to analyse these data for phylogenetic purposes. (Note, incidentally, that if I used a nucleotide sequence instead of an amino acid sequence as my example, then I would choose a sequence coding for a structural RNA rather than a protein.)

The four different objectives for sequence alignment are: (i) structure prediction, (ii) sequence comparison, (iii) database searching, and (iv) phylogenetic analysis. The alignment that is best suited to one of these purposes is not necessarily the one that is best suited to any of the other purposes. Let's consider each of these purposes in turn, using the pairwise alignment of 1qh5A and 1smlA for illustration.

The objective when using a sequence alignment for structure prediction is to deduce the secondary and tertiary structure of a gene product from knowledge of the gene sequence (reviewed by Gardner and Giegerich 2004; Simossis and Heringa 2004). That is, we use the gene sequence to predict the structure of a protein or RNA molecule based on alignment of the sequence to a gene (or genes) for which the structure is already known, and from which we can then infer other characteristics such as residue accessibility, functional specificity and tertiary interactions. For this procedure to be successful, we need to align those residues that occupy the same 3-dimensional position in the protein or RNA. This is a tricky business, but several computer programs exist to automate the procedure between pairs of molecules. In the example (Fig. 1*a*), I have used the data from the FSSP database of pairwise structure alignments (Holm and Sander 1996), which employs the DALI program. I have indicated the common structure along with the alignment, showing where the two structures are known to agree closely. The last three amino acids in the section of sequence shown are not aligned because the structures differ considerably at that point. However, the sequences are otherwise alignable, which is to be expected given that they are classified in the same protein-structure superfamily.

Sequence comparison [objective (ii)] is a rather diverse topic, and can include everything from: estimation of consensuses and genetic distances between sequences; the prediction and annotation of functional sites within sequences; gene prediction, identification and validation; primer and drug design; and on to the classification of protein

```
(a) Structure
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADS---LSA
1qh5A      TPCHTSGHICYFVSK-PGGSEPPAVFTGDTLF???
Structure --BBBTT-SSSSSS---TT----SSSS--TSS---

(b) Function
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1qh5A      TPCHTSGHICYFVSK-PGGSEPPAVFTGDTLF
Function  -A-ZAA----------------------A--A

(c) Database searching
1smlA      ??MAGHTPGSTAWTWTDT---RNGKPVRIAYADSLSA
1qh5A      TPCHTSGHICYFVSKPGGSEPPAVFTGDTLF??????
Consensus TPHPGHGPGHVVVYLGGG---KVLFTGDLLFSGGCGR

(d) Phylogeny
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1qh5A      TPCHTSGHICYFVSKPG-GSEPPAVFTGDTLF

(e) Global similarity
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1qh5A      TPCHTSGHICYFVSKPGGSEPPAVFTGDTLF?

(f) Local similarity
1smlA      ---magHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1qh5A     tpc---HTSGHICYFVSKPGGSEPPAVFTGDTLF?

(g) Structure
1smlA      M-AGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1k07A      T-PGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2       ?GKGHTEDNIVVWLPQ------YNILVGGCLVK
1jjtA      ?GPGHTPDNVVVWLPE------RKILFGGCFIK
1znbA      ?GGGHATDNIVVWLPT------ENILFGGCMLK
1qh5A      T-PCHTSGHICYFVSK-PGGSEPPAVFTGDTLF
Structure ----------SSSS----------SSSS-----

(h) Function
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1k07A      TPGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2       GKGHTEDNIVVWLPQ------YNILVGGCLVK
1jjtA      GPGHTPDNVVVWLPE------RKILFGGCFIK
1znbA      GGGHATDNIVVWLPT------ENILFGGCMLK
1qh5A      TPCHTSGHICYFVSK-PGGSEPPAVFTGDTLF
Function  -A-ZAA----------------------A--A

(i) Database searching
1smlA      MAGHTPGSTAWTWTDTRNG----KPVRIAYADSLSA
1k07A      ????TPGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2       ??GKGHTEDNIVVWL--------PQYNILVGGCLVK
1jjtA      ??GPGHTPDNVVVWL--------PERKILFGGCFIK
1znbA      ??GGGHATDNIVVWL--------PTENILFGGCMLK
1qh5A      ??TPCHTSGHICYFVSKPGG---SEPPAVFTGDTLF
Model      THHPGHGPGHVVVYL--------P-GKVLFTGDLLF

(j) Phylogeny
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1k07A      TPGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2       GKGHTEDNIVVWLPQ------YNILVGGCLVK
1jjtA      GPGHTPDNVVVWLPE------RKILFGGCFIK
1znbA      GGGHATDNIVVWLPT------ENILFGGCMLK
1qh5A      TPCHTSGHICYFVSKPG-GSEPPAVFTGDTLF

(k) Global similarity
1smlA      MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1k07A      TPGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2       GKGHTEDNIV-VWLPQYN-----ILVGGCLVK
1jjtA      GPGHTPDNVV-VWLPERK-----ILFGGCFIK
1znbA      GGGHATDNIV-VWLPTEN-----ILFGGCMLK
1qh5A      TPCHTSGHIC-YFVSKPGGSEPPAVFTGDTLF

(l) Local similarity
1smlA      MAGHTPGSTAWTWT-DTRNGKPVRIAYADSLSA
1k07A      TPGHTRGCTTWTmklKDHGKQYQAVIIGSigv-
3bc2       ?GKGHTEDNIVVWL-PQYNI----LVGGCLVK?
1jjtA      ?GPGHTPDNVVVWL-PERKI----LFGGCFIK?
1znbA      ?GGGHATDNIVVWL-PTENI----LFGGCMLK?
1qh5A      TPCHTSGHICYFVS-KPGGSEPPAVFTGDTLF?

(m) Progressive similarity
1smlA      ????????????????MAGHTP--GSTAWTWTDTRNGKPVRIAYADSLSA
1k07A      ??????????????????TPGHTR--GCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2       ?GKGHTEDNIVVWLPQYNILVG--GCLVK?????????????????----
1jjtA      ?GPGHTPDNVVVWLPERKILFG--GCFIK?---?????????????----
1znbA      ?GGGHATDNIVVWLPTENILFG--GCMLK?????????????????----
1qh5A      T-PCHTSGHICYFVSKPGGSEPPAVFTGDTLF?????????????????----

(n) Sequence-profile
1smlA        MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1k07A        TPGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2         GKGHTEDNIVVWLPQ------YNILVGGCLVK
1jjtA        GPGHTPDNVVVWLPE------RKILFGGCFIK
1znbA        GGGHATDNIVVWLPT------ENILFGGCMLK
1qh5A (seq)  TPCHTSGHICYFVSKPGGSEPPAVFTGDTLF?

(o) Profile-profile
1smlA (prof2) MAGHTPGSTAWTWTDTRNGKPVRIAYADSLSA
1k07A (prof2) TPGHTRGCTTWTMKLKDHGKQYQAVIIGSIGV
3bc2  (prof1) ?GKGHTEDNIVVWLPQ-----YNILVGGCLVK
1jjtA (prof1) ?GPGHTPDNVVVWLPE-----RKILFGGCFIK
1znbA (prof1) ?GGGHATDNIVVWLPT-----ENILFGGCMLK
1qh5A (prof1) -TPCHTSGHICYFVSKPGGSEPPAVFTGDTLF
```

Fig. 1. Alignments of a small section of the amino acid sequence of the metallo-β-lactamase protein-domain superfamily (Lactamase B) for six taxa. The alignments produced using various criteria and strategies are shown, thus illustrating just how different these alignments can be even for the same sequences. (*a*) Pairwise alignment from the FSSP database, showing the secondary structure: S = H-bonded β-strand; B = bend; T = H-bonded turn; – = random coil or equivocal structure. (*b*) Pairwise alignment based on information from Carfi *et al.* (1995), showing the functional sites: Z = direct zinc-binding; A = enzymic active site. (*c*) Pairwise alignment from searching the NCBI Conserved Domain Database, showing the consensus sequence used in the database. (*d*) Pairwise alignment produced manually from a consideration of structure, function and sequence similarity. (*e*) Pairwise alignment from the default settings of the MSA version 2.0 computer program. (*f*) Pairwise alignment from the default settings of the LAlign version 2.0u66 computer program, with lower case letters indicating unaligned residues. (*g*) Multiple alignment from the Homstrad database, showing the secondary structure: S = H-bonded β-strand; – = random coil or equivocal structure. (*h*) Multiple alignment based on information from Carfi *et al.* (1995), showing the functional sites: Z = direct zinc-binding; A = enzymic active site. (*i*) Multiple alignment from searching the Pfam database, showing the template used by the hidden markov model in the database. (*j*) Multiple alignment produced manually from a consideration of structure, function and sequence similarity. (*k*) Multiple alignment from the default settings of the MSA version 2.0 computer program. (*l*) Multiple alignment from the default settings of the DiAlign version 2.2.1 computer program, with lower case letters indicating unaligned residues. (*m*) Progressive multiple alignment from the default settings of the ClustalW version 1.83 computer program. (*n*) Sequence–profile alignment from the default settings of ClustalW. (*o*) Profile–profile alignment from the default settings of ClustalW. In all cases the full protein domain was aligned (163–192 amino acids), but only one small section of the alignment is shown, where ? = an amino acid from outside the section and – = a gap.

and RNA families (Miller 2001; Chain *et al*. 2003; del Sol Mesa *et al*. 2003). Therefore, the usual objective of sequence alignment for comparative purposes is to juxtapose residues representing conserved sequence features (e.g. conserved motifs, such as occur at active sites and binding sites, or signal sequences and transmembrane regions). This allows, for example, functional predictions for sequences of unknown function or detection of unsuspected functions in known sequences, because it is evolutionary constraints due to function that usually create the conserved sequence features. In the example (Fig. 1*b*) I have used the information from Carfi *et al*. (1995), who discuss the catalytic role of the sequence elements of zinc-binding lactamases. I have indicated the protein ligands that directly bind the Zn^{2+} in the functional protein, as well as those residues in the alignment that have a role in the catalytic (i.e. enzymatic) activity. Note that this involves aligning the final three amino acids in the section of sequence shown, as they perform the same functional roles. Otherwise, the function alignment is the same as the structure alignment because the structure is constrained by the function.

The objective when using a sequence alignment for database searching [objective (iii)] is to maximise the distinction between homologous and non-homologous sequences (reviewed by Pearson and Sierk 2005). That is, we want the high-scoring database matches to be sequences that are homologous to our query sequence, and non-homologous sequences to be low-scoring matches. To this end, we can simply search for a (nearly) perfectly matched sequence, or we can search for the appropriate family of homologous protein domains or RNA families. In the latter case, methods have been developed that use: 'fingerprints' consisting of small conserved motifs; regular expression patterns or consensus templates representing multi-sequence profiles; or hidden markov models expressing the positional probabilities. In the example (Fig. 1*c*) I have searched the Conserved Domain Database (Marchler-Bauer *et al*. 2005), which employs consensus templates, and in the figure I show the resulting alignment of each of the two sequences to the consensus template. The searches were successful, in the sense that they produced high-scoring matches only to the Lactamase B domain. However, it is obvious that the pairwise sequence alignment implied by the search results is very different from the previous two alignments.

For phylogenetic analysis, the objective of sequence alignment is to produce plausible hypotheses of evolutionary homology among the residues. That is, we hypothesise that each of the aligned residues has descended from a common ancestral residue. Note that this is distinct from database searching, which searches for homology between the genes as a whole rather than homology between the individual residues. Indeed, alignment for structural, comparison or phylogenetic purposes assumes that you have already determined that the sequences themselves are homologous. It is also distinct from structure and function prediction, where there is no necessity that evolutionarily homologous residues be aligned. Furthermore, phylogenetic alignment is a very different objective from the other three purposes in that it explicitly involves historical and therefore unobservable events. Whereas the success of the previous objectives is amenable to some experimental testing in all three cases, there is no gold standard of phylogeny with which we can compare an alignment, and so there is no straightforward means either to produce or to assess a phylogenetic alignment. This is discussed in more detail in a later section. In the example (Fig. 1*d*) I have, as my criteria, used structural and functional similarity plus sequence similarity at both the amino acid and nucleotide levels. The alignment is thus basically the same as the functional alignment but with several of the ambiguous residues re-aligned to increase the sequence similarity at the nucleotide level.

Thus, we have four different pairwise sequence alignments, each optimised for a different purpose. Unfortunately, in practice the most commonly used automated alignment procedures use little more information than the overall phenetic similarity between the sequences. That is, some scoring scheme is employed (a function encoding the objective) that measures the degree of similarity between pairs of residues, and the alignment procedure then tries to optimise the overall score for the sequence of juxtaposed residues (e.g. to maximise the sum of the scores produced by the string of aligned residue pairs). Hence, the complexity of the four alignment purposes is reduced to finding the alignment with the maximum similarity among the sequences, so that similarity is used as a substitute for each of the four criteria outlined above. This similarity alignment is often referred to as 'the optimal' alignment, but it really should be 'one of the optimal' alignments.

In this context, there are two main optimisation options: to use either global or local similarity. In the former case, it is the total score for the entire alignment that is optimised, whereas in the latter case some residues are allowed to be unaligned and thus do not contribute to the score. In a global alignment all residues are paired and scored, and there is a trade-off along the alignment, where potentially high-scoring pairs do not align because an even higher score can be achieved with a mutually exclusive arrangement. A local alignment is a single block of high-scoring residue pairs, which does not necessarily include the whole sequence. If a series of local alignments is applied to the sequences, then only small ungapped blocks or equivalent segments will be displayed (e.g. by database-search or motif-recognition software), interspersed with non-scoring residues.

In the example, the global alignment (Fig. 1*e*) was produced by the MSA computer program (Gupta *et al*. 1995) and the local alignment (Fig. 1*f*) by the LAlign program (Huang and Miller 1991). Note that the global alignment does not insert a gap in the middle of the 1qh5A sequence, as do

all of the previous alignments. This allows both a proline (P) and an aspartic acid (D) to be paired, thus increasing the overall pairwise similarity. The local alignment aligns these same pairs, but it leaves unaligned the first three amino acids in the section of sequence shown. Neither of these similarity alignments matches the best alignments determined for any of the four purposes outlined above. In this case, similarity is not a perfect substitute for any of the other four criteria.

The four different modes of sequence alignment are: (i) pairwise alignment, (ii) multiple alignment, (iii) sequence–profile alignment, and (iv) profile–profile alignment. (Note that there are many ways of classifying the modes of sequence alignment. I am focusing here on the point of view of the user rather than the developer of the methods. Also, I am not including manual alignment.) All of the alignments discussed so far have been pairwise alignments, since only two sequences were involved. I will illustrate each of the other three modes by adding four more sequences to the existing pair. One of these new sequences (1k07A) is relatively similar to the 1smlA sequence, while the other three are similar to each other but quite different from both 1smlA and 1qh5A.

The basic distinction between multiple and pairwise alignment is that in the former situation each pair of sequences cannot necessarily be optimised, but instead there are tradeoffs where some pairs may have suboptimal alignments in order to increase the overall optimality among the group of sequences. That is, the pairwise alignments are judged simultaneously in the context of the other sequences, so that the pairwise alignments occur within a larger framework rather than in isolation. This distinction can be illustrated by comparing the pairwise alignments discussed above (Fig. 1*a–d*) with how these same two sequences are aligned within multiple alignments designed for the same four purposes (Fig. 1*g–j*).

For the structure alignment (Fig. 1*g*) I have used the data from the Homstrad database of multiple structure alignments (Stebbings and Mizuguchi 2004), which has manually curated alignments based originally on the STAMP program. Note that the last three amino acids in the section of sequence shown are now aligned, but otherwise the pairwise structure alignment is the same as before (Fig. 1*a*). This difference is a result of the changed context within which the structural super-positions are assessed. Indeed, aligning sequence A against structure B does not necessarily produce the same result as aligning sequence B against structure A. For the sequence-comparison alignment (Fig. 1*h*) I have used the function information from Carfi *et al*. (1995). As expected, the pairwise alignment is the same as before (Fig. 1*b*), since the functions of the residues are the same. For the database-search alignment (Fig. 1*i*) I have searched the Pfam database (Finn *et al*. 2006), which employs hidden markov models, and in the figure I show the resulting alignment of each of the six sequences to the model template. All of the searches were successful, in the sense that they produced high-scoring matches only to the Lactamase B domain. The pairwise sequence alignment implied by the search results is quite different from the previous search alignment (Fig. 1*c*), now being more similar to the other alignments. This is often the case when using markov models, which are considered to be superior to the use of consensus sequences, although they are much slower to calculate. For the phylogenetic alignment (Fig. 1*j*) I used the same principles as outlined above. As expected, the pairwise alignment is the same as before (Fig. 1*d*), since the alignment represents a set of hypotheses of ancestry, which should not change just because other descendants are included in the alignment.

For multiple alignment in practice, the most commonly used automated alignment procedures also use either global or local similarity as their criterion, rather than optimising the alignment for any of the four purposes that I have outlined. In the example, the global alignment (Fig. 1*k*) was produced by the MSA program and the local alignment (Fig. 1*l*) by the DiAlign2 program (Morgenstern 1999). In both cases the pairwise alignment between 1smlA and 1qh5A differs from before (Fig. 1*e*, *f*), because there are trade-offs in the assessment of similarity. Moreover, not only do these two multiple alignments differ from each other they also differ from all of the other multiple alignments. So, once again, similarity is not a perfect substitute for any of the other four criteria.

Even more importantly, the most commonly used automated procedures for multiple sequence alignment do not actually optimise similarity but instead use heuristic procedures to approximate the optimal result. The best-known heuristic technique is progressive alignment, where the sequences are aligned in some order rather than simultaneously. That is, there is no attempt at a simultaneous assessment of all possible pairwise alignments, but instead the pairwise assessments occur sequentially. For the example, I have used the most popular sequence-alignment program, ClustalW (Thompson *et al*. 1994). Clearly, the resulting alignment (Fig. 1*m*) bears little relationship to any of the other alignments. In this case, the procedure has aligned the group of three similar sequences reasonably well, along with the pair of similar sequences, but it has not successfully aligned these two groups to each other or to the sixth sequence. This problem arises from the low levels of pairwise identity among the sequences, as the twilight zone is not a region where the progressive-similarity strategy can be expected to work well (as discussed in a later section).

Sequence–profile alignment is a third mode of alignment, where a single sequence is aligned against a pre-existing multiple alignment (which is then called a profile; Wang and Dunbrack 2004). The profiles do not have to be actual alignments, but may be ‘condensed’ alignments such as consensus templates or hidden markov models. Either

way, the aligned residues in the profile remain aligned, although gaps can be inserted between aligned positions, as well as within the other sequence. For the example, I have aligned the 1qh5A sequence (the least similar sequence) against the profile formed from the phylogenetic multiple alignment of the other five sequences (Fig. 1*n*), using the ClustalW program. Note that the alignment of 1qh5A and 1smlA is similar to that shown for the global-similarity pairwise alignment (Fig. 1*e*) rather than that for the progressive-similarity alignment (Fig. 1*m*). This is because the extra information in the profile is used for assessment (i.e. the profile already has a simultaneous assessment of the sequences that it contains), so that sequence–profile alignments are expected to be superior to progressive alignments (Edgar and Sjölander 2004).

Profile–profile alignments involve, as the name suggests, the alignment of two profiles. For the example, I have aligned the profile formed from the phylogenetic multiple alignment of the 1smlA and 1k07A sequences (the two badly aligned sequences in the progressive alignment) against the profile formed from the phylogenetic multiple alignment of the other four sequences (Fig. 1*o*), using the ClustalW program. The pairwise alignment of 1qh5A and 1smlA is improved compared to the progressive-similarity alignment (Fig. 1*m*), because the extra information in the two profiles is used for assessment, so that profile–profile alignments are expected to be superior even to sequence–profile alignments (Ohlson *et al*. 2004).

So, my conceptual framework for sequence alignment is this: there are four objectives (structure prediction, database searching, sequence comparison, phylogenetic analysis) and there are four different modes (pairwise, multiple, sequence–profile, profile–profile), but in practice we don't actually optimise these methods based on these criteria, but instead we optimise either global or local similarity for pairwise alignment and use progressive similarity for the other three. This complexity should make it clear why alignment is considered to be problematic. This complexity will usually not be apparent for closely related (and thus similar) sequences, but as sequences diverge over evolutionary time the different objectives, modes and practices will result in different alignments.

Finally, it is important to note that most of the sequence-alignment computer programs were developed originally for sequence comparison, and they have been applied subsequently to the other three purposes in an *ad hoc* manner, without regard for their suitability. Specialist programs have recently been developed for structure alignment, such as DALI, Mammoth and VAST for pairwise amino-acid alignments, CE, SSAP and STAMP for multiple amino-acid alignments, and RNAforester for RNA alignments. Nevertheless, alignment errors are still considered to be the biggest problem in structure prediction (Cozzetto and Tramontano 2005; Kolodny *et al*. 2005). There are also now specialist programs for database-search alignment, such as the pairwise programs SSearch, FASTA and BLAST, and profile-search programs such as PSI-BLAST, Compass and COACH. In addition, the search for functionally conserved subsequences (i.e. sequence comparison) has focused on the use of local alignment strategies, with programs such as Consensus, MEME and BioProspector. Here, the major limitation seems to be the weakness (i.e. subtlety) of the motifs rather than the alignment strategy used (Frith *et al*. 2004).

However, there can be no such thing as a computer program for phylogenetic alignment since this involves a study of historical accidents, for which there can be no objective function to optimise. Homology is an inference, not an observation (unless we have a time machine), and hence we cannot expect to optimise an alignment with respect to homology. Since phylogenetic alignment is not amenable to an automated (i.e. computerised) procedure, no such 'recent developments' have occurred in this field. Therefore, most people still seem to use the original global sequence-comparison programs. This has long been known to have a negative effect on phylogenetic analyses if the alignments are unsuitable (Ellis and Morrison 1995; Morrison and Ellis 1997). The problems caused by this issue are the topic of this review.

Most of the recent developments in the commonly used alignment programs have also been designed for sequence comparison, especially for amino acid sequences rather than nucleotide sequences. This is mainly because the larger 'alphabet' of amino acids (20) compared with nucleotides (4) allows more information to be used in making alignment decisions, although structural modelling of proteins has also motivated some of the improvements. (Note that the codon alphabet of 61 should be even better, but it has so far been too unwieldy to be put into practice easily.) In addition, database searches are usually more effective when using amino acid sequences compared with nucleotide sequences, and this has provided impetus for other improvements. However, for nucleotide alignment, which is the prevalent mode in phylogenetic analyses, very little has changed in practice over the last 20 years (Taylor 1996; Phillips *et al*. 2000; Raghava *et al*. 2003), when the progressive-alignment strategy was first developed (Hogeweg and Hesper 1984; Feng and Doolittle 1987; Taylor 1987). Nevertheless, some of these developments could usefully be moved across to phylogenetic analysis of nucleotides as well, and they are discussed in more detail in several sections below.

Homology and similarity

Characters and character-state homology

For phylogenetic analysis, homologous rather than analogous characters and character states must be compared across the taxa. That is, for all of the taxa we must be comparing like

with like regarding the evolutionary origin of the attributes. While the term homology has been used historically to refer to a wide variety of concepts (Wagner 1989; Sluys 1996; Butler and Saidel 2000), the evolutionary concept of homology refers to the relationships of features that are shared among taxa due to common ancestry (i.e. they all inherited the feature from their most recent common ancestor). Systematists and phylogeneticists have long insisted on this definition, and there have been calls for all molecular biologists to use it as well (Reeck *et al*. 1987).

In phylogenetic analysis this definition serves the very useful purpose of highlighting the fact that similarity ≠ homology. In phylogenetics (and comparative biology in general) similarity = homology + analogy, instead. Analogy refers to similarity resulting from the same function rather than similarity resulting from the same evolutionary origin. Analogy will lead to incongruences among the characters compared to the relationships shown by homology, and these will confound our ability to detect homology. That is, some of the apparent relationships among taxa will be due to homology and some will be due to analogy, and these two patterns of relationship are unlikely to be in agreement. We are then caught in the bind of trying to disentangle the two patterns, because the one due to homology is the one that we really want, in an evolutionary context. So, mistaken hypotheses of homology are the primary source of error in evolutionary studies.

Consequently, one of the essential steps in the cladistic reconstruction of phylogenetic history is the establishment of hypotheses of character and character-state homology among the taxa being studied. The choice of characters to be included in a phylogenetic analysis may be somewhat arbitrary, but can include intrinsic (phenotypic or genotypic) attributes from morphology, anatomy, embryology, behaviour, physiology, ultrastructure, cytology, biochemistry, and immunology. Each of these disciplines may have their own criteria of homology (Sluys 1996), but the important points are that the characters used in the analysis are hypothesised to reflect the evolutionary history of the taxa, and that the character states of a single character are hypothesised to have a unique evolutionary origin. Note that these are hypotheses not observations; we cannot observe homologies in nature but must instead speculate about their existence. We can *observe* similarity, from which we then *infer* either homology or analogy. This is what makes sequence alignment so difficult—it is not intrinsically an empirical subject, and yet we must make it so to the best of our ability. We cannot know where the evidence is that will reveal the evolutionary history of the sequences—we can only hypothesise that there will be some evidence somewhere in the sequences. This is a radical difference from experimental science, where our experimental hypothesis specifies exactly where to look for the relevant evidence.

Alignment of molecular sequences for phylogenetic purposes is thus a series of hypotheses of homology among the taxa, with one hypothesis of homology for each position (nucleotide or amino acid) in the aligned sequences. That is, we hypothesise that the nucleotides or amino acids at each position are descended from the same positional residue in a common ancestral sequence. 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). Differences in residues at an aligned position thus represent explicit hypotheses about the evolutionary events that caused the differences, and our hypotheses about these events should be plausible and parsimonious. Plausibility is an obvious requirement for any hypothesis, while parsimony is simply the methodological convention that we should not create hypotheses that are more complex than is strictly necessary. (Note that this is descriptive parsimony, where we prefer simpler explanations, rather than ontological parsimony, where we claim that evolution itself necessarily acts parsimoniously; Johnson 1982.)

The idea that alignments represent hypotheses about evolutionary events can be made clear by a simple example (Fig. 2). The sequences are from the study by O'Donnell *et al*. (2000) of the phylogeography of the microfungi causing wheat scab and blight. The short stretch of sequence shown represents, as far as evolutionary events are concerned, a series of dinucleotide (TC) repeats with subsequent substitutions in the second nucleotide. The aligned Cs and

```
             1       10        20        30        40
NRRL_28338   TCAACTAACCGTGA|TCTCTCTC----|CAGGCATTATTGGT
NRRL_28334   TCAACTAACCGTGA|TCTTTCTCTC--|CAGGCATTATTGGT
NRRL_28065   TCAACTAACCGTGA|TCTTTCTCTC--|CAGGCATTATTGGT
NRRL_28062   TCAACTAACCGTGA|TCTCTCTC----|CAGGCATTATTGGT
NRRL_13393   TCAACTAACCGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_25805   TCAACTAACCGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_13818   TCAACTAACCGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_6101    TCAACTAACCGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_26755   TCAACTAACCGTGA|TCTCTCTCTCTA|CAGGCATTATTGGT
NRRL_26754   TCAACTAACCGTGA|TCTCTCTCTCTA|CAGGCATTATTGGT
NRRL_26752   TCAACTAACCGTGA|TCTCTCTCTCTA|CAGGCATTATTGGT
NRRL_29105   TCAACTAACCGTGA|TCTCTCT-T---|CAGGCATTATTGGT
NRRL_29011   TCAACTAACCGTGA|TCTCTCT-T---|CAGGCATTATTGGT
NRRL_26916   TCAACTAACCGTGA|TCTCTCT-T---|CAGGCATTATTGGT
NRRL_29020   TCAACTAACCGTGA|TCTCTCT-T---|CAGGCATTATTGGT
NRRL_29148   TCAACTAACCGTGA|TCTCTCTCTC--|CAGGCATTATCGGT
NRRL_28718   TCAACTAACTGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_28585   TCAACTAACTGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_2903    TCAACTAACTGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
NRRL_29010   TCAACTAACTGTGA|TCTCTCTCTC--|CAGGCATTATTGGT
```

Fig. 2. Alignment of nucleotide sequences from *Fusarium pseudograminearum* (blight) strains, illustrating the concept that aligned columns represent explicit hypotheses about the evolutionary events that created the sequence patterns. The gene is labelled by O'Donnell *et al*. (2000) as 'similar to ammonia ligase 1'. The multiple alignment was produced by the ClustalW 1.83 program, with default settings. The vertical bars delimit a region of dinucleotide repeats.

Ts at position 18 are thus hypothesised to have arisen from a repeat followed by a substitution, as are the aligned As at position 26. However, the four T–T– at positions 21–24 are not treated that way—instead the alignment actually treats them as having arisen from two separate repeats followed by two deletions. That is, by aligning the Ts we are explicitly claiming that they are homologous with the other Ts, and so they arose from the same duplication events. If, on the other hand, the second Ts were to be moved one position to the left (i.e. TT—) then the evolutionary hypothesis would also involve one repeat followed by a substitution. This is a more parsimonious evolutionary scenario (involving two events instead of four), and thus it represents a more plausible set of hypotheses of positional homology among the residues. An alternative scenario is that the second set of Ts is not part of a dinucleotide repeat but a separate insertion, so that the evolutionary hypothesis involves one repeat followed by a deletion and an insertion (i.e. three events). For this scenario the second set of Ts should be aligned against a gap in all of the other sequences.

Homology assessment can be considered to involve two steps (de Pinna 1991). The first step is the conjecture, before 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 before phylogenetic analysis (an assessment of 'essential samenesss' without reference to an ancestor) while secondary homology is a corroborated homology assessment subsequent to the analysis (an assessment of 'congruence' that explains the sameness as resulting from common ancestry). From this perspective, sequence alignment is primary homology assessment (Mindell 1991; Brower and Schawaroch 1996; Phillips *et al*. 2000; Phillips 2006). Also, secondary homology does not mean that the homology assessment is necessarily correct, since errors may also be congruent.

It has been traditional in phylogenetic analyses to keep assessment of primary and secondary homology separate. For example, it is usually considered necessary to have specialist expertise when assessing morphological, anatomical or ultrastructural characters (or the sorts of macro-biological characteristics often used for unicellular organisms, such as host, tissue and vector specificities), and researchers will spend hours contemplating the rationale for their decisions regarding primary homology assessment. Indeed, as far as time is concerned the assessment of these phenotypic characters may be the major part of any one study, and it may take up most of the space in the subsequent publication. Secondary homology assessment, on the other hand, is a process requiring a common expertise from all phylogeneticists. In this sense, sequence alignment (primary homology assessment) is separate from tree building (secondary homology assessment).

There are two basic concepts within primary homology (Brower and Schawaroch 1996): topographic identity and character-state identity. The first of these refers to the identification of comparable features among the taxa concerned (i.e. the creation of a blank characters × taxa data matrix), while the second refers to the decision about which character-states are to be classified as identical (i.e. the filling in of the cells of the data matrix) (Brower and Schawaroch 1996). As discussed below, the distinction between topographic identity and character-state identity is important when considering the relationship between molecular data and phenotypic data. Furthermore, it is useful to remember that homology is a hierarchical concept (Rieppel 1994), so that characters may be homologous at a more general (inclusive) level but not at a more specific level (e.g. bird wings and bat wings are homologous as forelimbs, which are common to all vertebrates, but not as wings, which arose independently in birds and bats). From this point of view, characters are just hypotheses of homology at a more inclusive level than those of character states (Patterson 1988).

When dealing with phenotypic data (e.g. morphology, anatomy), characters and their states can be postulated as homologous on the basis of their structural, positional, ontogenetic, compositional and / or functional correspondences; and they can be postulated between different taxa so as to maximise the number of one-to-one correspondences of their parts. That is, the features are decomposed into their constituent parts, and these are compared in terms of their positional and connectional relationships (i.e. topology) (Rieppel 1994). This concept may be problematic (for example, deciding on what constitutes the ultimate parts to be compared), but it can be put into practice through a detailed comparative and / or experimental study of the organisms concerned (what has traditionally been called comparative anatomy), where the constancy of the topological relationships is used as the main criterion for recognising primary homology (Rieppel 1994). In particular, structural and positional similarity of *complex* structures has traditionally been taken as good evidence of primary homology (Donoghue and Sanderson 1994; Rieppel and Kearney 2002). So, for phenotypes determining topographic identity may, in fact, be uncontroversial (de Pinna 1991), while character-state identity may be more complicated (Brower and Schawaroch 1996).

Molecular homology

When dealing with molecular data concepts of homology have often been rather confused (Winter *et al*. 1968; Reeck *et al*. 1987; Patterson 1988; Hillis 1994; Fitch 2000;

De Laet 2005), with the word homology being used to mean several unrelated things, which could perhaps better be given alternative names. In particular, 'sequence homology' is often used as a synonym for 'sequence identity' (i.e. the number of nucleotides or amino acids that are inferred to be held in common between two sequences). These are not necessarily the same thing (Reeck *et al*. 1987), since similarity can be the result either of common ancestry or of chance convergence, parallelism or reversal; and 'isology' may be a better term to use (Wegnez 1987).

Nevertheless, for primary homology at a general level, considerable thought has been given to the study of molecular data (Patterson 1988; Williams 1993). For example, it has long been recognised that the sequences being compared must themselves be homologous rather than analogous. However, for analyses using molecular sequence data, the assessment of primary homology also involves the alignment of the nucleotides or amino acids; that is, 'positional homology' for the components of the homologous sequences. This point has not received very much consideration in discussions of molecular homology to date (Dewey and Pachter 2006; Phillips 2006), as these discussions have mainly focused on the more general level of homology among genes (e.g. Patterson 1988; Hillis 1994). Nevertheless, positional homology is just as important for the assessment of character and character-state homology (*contra* Wheeler 2005).

The positional homologues can be represented by either identical character-states (nucleotides or amino acids) in all of the sequences, substitutions in one or more of the sequences (representing point mutations), or insertions / deletions (indels) in one or more of the sequences—substitutions and indels are produced by different mechanisms and so are evolutionarily distinct. The concepts of primary homology in molecular and morphological studies are thus fundamentally the same (Patterson 1988; de Pinna 1991; Williams 1993), as sequence alignment is simply the process of determining topographic identity (Brower and Schawaroch 1996). Indeed, the recognition of the constituent parts that are to be compared (i.e. the nucleotides and amino acids) is usually considered to be unproblematic for molecular data, compared with the problems encountered for phenotypic data.

Unfortunately, the concept that positional alignments should explicitly reflect evolutionary homology has often been ignored in molecular biology. In‘fact, most computer-based alignment methods use phenetic pattern-matching algorithms (discussed in the next section), and their procedures are thus based on maximising sequence similarity (i.e. isology). This is a result of the idea that the dominant test for homology in molecular studies concerns empirical observations of similarity — that is, homology in classical and molecular biology *do* differ in that assessments of similarity are a strong test of homology for molecular data but are not for phenotypic data (Patterson 1988). This has been seen as an equivalent of the attempt to maximise the number of one-to-one correspondences of phenotypic features. However, this idea follows from the consideration of molecular sequences as being one-dimensional (i.e. a string of nucleotides or amino acids), rather than being three-dimensional in the same way as are phenotypic features, so that the recognition of homology (as opposed to its definition) is seen as being simply a statistical problem of similarity assessment (Aboitiz 1987; Patterson 1988). However, this view is incorrect when the focus is shifted from the level of gene homology to the level of positional homology (Donoghue and Sanderson 1994).

Note that assessment of character-state identity is often assumed to be uncontroversial for sequence data. That is, an adenine is obviously an 'adenine' and a proline is obviously a 'proline'. However, this misses the essential point that, when assessing homology, identity of character-states needs to reflect evolutionary history. Therefore, two adenines represent the same character state *only* if they originated as adenines as a result of the same evolutionary event (i.e. an adenine is only the same as an adenine with which it is aligned). That is, character states can only be defined by reference to both nucleotide type and to position, and any sequence alignment procedure proceeds by shuffling character states among characters (which usually does not happen when dealing with morphological characters, for example). Furthermore, if multiple substitutions have occurred then the apparent similarity of any two adenines may represent homoplasy rather than homology. Assessing character-state homology is the same as assessing character homology but just at a different level of generality.

Structure, function and homology

For molecular data it has been traditionally accepted that there is little possibility of further investigations (similar to comparative anatomy) to assess the topographic identity of alignments (Patterson 1988), and this is why the phenetic algorithms have been employed in computerised sequence alignment. So, in practice primary homology assessment has been very different in molecular analyses compared with analyses of phenotype. It is important to recognise this, because the correct formulation of hypotheses of homology is just as important for molecular data as it is for phenotypic data.

Nevertheless, I maintain that it *is* possible to employ detailed structural and positional analyses of molecular sequences to assist in the alignment of homologous positions, just as it is possible to use them for the analysis of phenotypic characters. This is because the sequences are, in most cases, merely the code for the production of a molecule (e.g. a protein or an RNA) in which certain active sites must be maintained, and consequently the order of the nucleotide or amino acid sequence is constrained by the structure and / or function of its end product. Comparative sequence analysis (Gutell *et al*. 2002; Errami *et al*. 2003) has shown that during

evolution the structure of the end product molecule has been better conserved than has the corresponding DNA or amino acid sequence, and therefore even quite divergent sequences must still produce a molecule with the same three-dimensional structure and the same molecular function. So, the primary structure of the sequences can continue to evolve, but only within the constraints needed to maintain the same secondary and tertiary structure (and therefore molecular function) of the molecule.

There are thus actually three components that can be used to help decide whether two sequences (and their parts) have evolved from a common precursor: (i) the DNA, RNA and / or amino acid sequences are similar; (ii) the three-dimensional structures of the product being coded for (e.g. RNA or protein) are similar; and (iii) the active sites and functional activities (e.g. catalytic mechanisms, enzyme-substrate interactions) are similar. Note that these were the three criteria that I used to formulate the 'phylogenetic alignments' in Fig. 1.

These three are related concepts but they are not identical, as structural, functional and sequence similarity may be mutually inconsistent (Shakhnovich 2005). For example, proteins are known that share a common structure and function but have a low sequence similarity (e.g. histone deacetylase and arginase), while others share sequence similarity and function but not structure (e.g. Bir A domain II and SH2 domain), others share similarity and structure but not function (e.g. lysozyme and α-lactaburin), and others have sequence similarity but share neither structure nor function (e.g. β-heamoglobin and cellulase E2). Thus, similarity of topology and shared functional constraints represent evidence on which to base hypotheses of homology, but they may still actually represent homoplasy rather than homology (for further examples see Morris and Cobabe 1991; Graham *et al*. 2000; Pearson and Sierk 2005).

Many recent advances in sequence alignment involve trying to explicitly incorporate (ii), and to some extent (iii), into the process. Thus, the sequence–structure–function relationship provides a mechanism for incorporating studies of primary homology into molecular sequence alignment (e.g. Kjer 1995, 2004; Hickson *et al*. 1996; Jennings *et al*. 2001; Simossis and Heringa 2004). As an example of the perceived importance of this relationship, evaluation of the quality of procedures for multiple sequence alignment is now frequently assessed using structure-based reference alignments as the best estimate of the 'correct' alignment, including BAliBASE (Thompson *et al*. 1999*a*, 2005; Bahr *et al*. 2001), OXBench (Raghava *et al*. 2003), PREFAB (Edgar 2004*b*), SABmark (Van Walle *et al*. 2005), IRMBase (Subramanian *et al*. 2005) and BRAliBASE (Gardner *et al*. 2005); and a similar approach has been taken to evaluation of pairwise alignment for database searches and structure prediction (Brenner *et al*. 1998; Domingues *et al*. 2000; Sauder *et al*. 2000; Marsden and Abagyan 2004; Marti-Renom *et al.* 2004). Note that this approach explicitly uses a biological criterion to judge the quality of the alignments, rather than whatever mathematical criterion was used to produce the alignments. It does not, however, make any explicit claim that a structural alignment must necessarily correspond to the true evolutionary alignment, since knowledge of this 'gold standard' would require a time machine.

The important conclusion for sequence alignment is that if the two- or three-dimensional structure of a molecule is known, even approximately, then the sequence alignment process can be constrained by that model. From this point of view, molecular sequences *are* three-dimensional in the same way as are phenotypic features (*contra* Patterson 1988), and adding dimensions to the assessment of primary homology must make these assessments more reliable (Donoghue and Sanderson 1994). That is, if topology is considered to be 'the ultimate operational clue to homology' (Rieppel 1994) for phenotypic characters, then the molecular equivalent must involve the secondary and tertiary structures of the molecules. Furthermore, these structures make explicit the frame of reference within which the topological relationships are to be assessed, which is quite problematic for phenotypic data.

Detailed structural and functional analyses thus have exactly the same role to play in homology assessments as they do for phenotypic data—they can provide *evidence* for primary homology (since we can never know about true evolutionary homology). This idea is not new, as it is implicit in Winter *et al*.'s (1968) definition of homology as '*structural* similarity among proteins greater than might be anticipated by chance alone' [my italics], but the potential role that structural considerations have to play in sequence alignment has generally been ignored to date. Indeed, these have often been considered as secondary to sequence similarity as criteria for alignment (e.g. Hillis 1994). Note that this viewpoint suggests an explicitly biological solution (the consideration of the biological relationship between structure and function) to the problem of finding the correct alignment, rather than the traditional purely mathematical solution (the search for an optimal alignment).

Gaps and recombination events

If we think of homologies as being character states shared by (at least) two organisms that are derived from a single transformation / mutation event in their common ancestor (Vogt 2002), then there are several complicating factors when trying to assess positional homology of nucleotides and amino acids (i.e. we have difficulties inferring the evolutionary events). These are examples of general problems when dealing with homology, but it is important to emphasise them here, as they will recur in later sections.

First, homologies are hierarchical, so that homologous structures at one spatial or temporal scale (i.e. level of

generality) are not necessarily homologous at some other scale. That is, positional homology in an alignment implies that the sequences are homologous, but the converse is not necessarily true. This means that any process that involves sequence transposition will lead to complex structural homology (Bledsoe and Sheldon 1990).

At the simplest level, two different aligned nucleotides are homologous if the evolutionary event that created their difference was a point mutation leading to a substitution. However, recombination, gene conversion and horizontal (or lateral) gene transfer create gene sequences that are spatial mosaics. In this context, aligned positions may not differ in a substitution event but instead differ in a recombination or transfer event involving a whole block of positions. So, at one level of generality the aligned nucleotides are not homologous (because their similarities and differences are not the result of ancestry by descent), but at a more general level they are homologous if the same piece of gene is involved in the event (i.e. the nucleotides in the genes do have a common ancestor further back in history). In this sense, horizontal gene transfer can be thought of as 'non-homologous recombination', since it involves a genomic region that is new to a genome, whereas 'homologous recombination' involves a new variant of a genomic region that is already present.

A similar situation occurs with regard to protein domains. Protein chains are made up of one or more domains, which are the basic structural and functional units, the combination of domains determining the function of the protein. A domain is thus a structurally and functionally defined region of a protein chain (which is self-stabilising and usually folds independently of the rest of the chain). Domains are normally not unique to the protein products of one gene, but instead occur in a variety of proteins and thus form domain families. Therefore, a single protein may not be composed of domains with a common ancestry by descent, while apparently unrelated proteins may actually be related by descent if they have a related domain. Alignment of positions within and between domains may thus be complex, as any two gene sequences may not be related over their entire lengths. This all means that only homologous positions at the relevant hierarchical level should be aligned. An alignment thus depends on the context, where an alignment may represent homology at one level of generality but not at a lower level.

The second complication concerns the mosaic nature of sequences created by the various events that lead to gaps being inserted in one or more of the sequences. That is, there are various ways in which a gap can arise, including long or short indels and sequencing mistakes. Whether a gap can be considered to be homologous to the residues at the same alignment position depends on the nature of the event that created the gap. Note that a distinction will be made here between gaps, which relate to sequences (the 'rows'), and evolutionary events, which relate to alignment positions (the 'columns').

On the one hand, if a gap arises from a single nucleotide insertion or deletion at an aligned position then it is homologous with the residues at that position, because the gap represents a single (inferred) evolutionary event just like a substitution. Thus, the gap could be treated as a fifth character state in a nucleotide alignment or a 21st state in an amino acid alignment. However, note that this is artificial, because a gap is not an observation, but rather we infer the gap after having observed the presence of a residue in at least one other sequence (Geiger 2002).

On the other hand, a multi-position insertion or deletion is not homologous with the aligned residues because there is no evolutionary event associated with each single gap position, so that the residues do not have an homologous character state at the gap. Instead, the relevant evolutionary event is the insertion or deletion of the block of residues, so that the gap is homologous to a set of residues in the ungapped sequences rather than to a single residue. In this sense, for substitutions each aligned position is a character and the residues are the character states, but for long indels each character covers more than one aligned position and the character states are presence / absence. One can even argue that for an insertion the gap does not represent the event at all but rather the absence of the event, so that the gap should be coded as 'inapplicable' rather than as a character state. In contrast, for a deletion the gap does represent the event, but the whole gap is a single character state.

Moreover, each indel event occurs on a particular branch of the evolutionary tree and therefore refers only to a subset of the sequences. So, if multiple indels occur in the same length of sequence, then some of the aligned gaps may represent 'inapplicable' in one subset of sequences but 'deletion' in another subset. This complexity is illustrated in Fig. 3.

Nevertheless, an evolutionary event (or more than one) has occurred for every gap, so this information should be

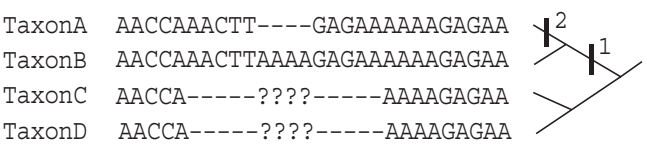

Fig. 3. Artificial nucleotide sequences for four taxa along with their phylogenetic history (rooted at the right), showing the complexity of homology that can be associated with gaps. After divergence from their common ancestor, the ancestor of TaxonA and TaxonB acquired an insertion (event 1), which is indicated by the gap in the sequences of TaxonC and TaxonD. The gap thus represents absence of the evolutionary event. TaxonA then underwent a deletion within the inserted sequence (event 2), which is indicated by a gap in TaxonA. This gap thus represents presence of the evolutionary event. Furthermore, this second event is inapplicable to TaxonC and TaxonD, which do not have the insertion, and this is indicated by missing characters in their sequences.

represented in the phylogenetic analysis somehow. Coding the gap as 'missing' is just a roundabout way of trying to leave the substitution information in the alignment without actually acknowledging the existence of the evolutionary event that created the gap (Geiger 2002). This leads to the concept of coding gaps as separate characters (i.e. presence or absence of a specified gap). If a gap represents a sequencing error, then it should be coded as 'missing', of course.

The third complication is simply an extension of the second one: at the molecular level there are actually several different mutational events that affect blocks of sequence simultaneously (Phillips 2006). These include (Benson 1997): inversions (replacement of a subsequence by its reversed sequence); translocations (removal of a subsequence and its insertion at another location); transpositions (copying of a subsequence to another location); and tandem duplications (copying a subsequence to an immediately adjacent position). All of these will confound assessments of positional homology, as it is the subsequences that are homologous but not necessarily the individual positions. That is, we can no longer treat individual sequence positions as independent characters. Such events can be expected to occur in longer sequences.

In summary, there are at least six processes that can create sequence differences that require alignment: (1) substitutions (which don't change the sequence length); (2) short insertions (such as microsatellite repeats or hairpin inversions); (3) short deletions; (4) long insertions (such as horizontal gene transfer); (5) long deletions (such as deletion of a helix); and (6) long replacements (such as recombination, conversion, inversion; these also do not change the length). It seems unlikely that we will ever be able to put all of these processes into a single quantitative model that could be used to guide sequence alignment. The computer programs to be discussed in the next section only work well for protein-coding sequences, where indels and replacements are relatively rare and can be modelled as different types of 'substitutions'. Indels are quite common in RNA-coding sequences and especially in non-coding sequences, making their alignment much more problematic.

Computerised sequence alignment

Pattern-matching alignment

If the goal of sequence alignment is to infer the true evolutionary relationships between sequences without knowledge of the evolutionary events themselves (Waterman 1995), then in practice the establishment of a sequence alignment requires a set of criteria to determine the 'success' of the alignment. To this end, there is an almost universal use of computerised pattern-matching algorithms, which produce so-called 'optimal' sequence alignments based on an assessment of phenetic similarity, as these methods are seen as providing 'explicit and objective rules for inferences of positional homology' (Hillis 1994). However, because homology $\neq$ similarity, there is no *necessary* reason to expect these algorithms to produce multiple-sequence alignments based on evolutionary homology (features shared due to common ancestry) as opposed to homoplasy (features shared due to chance similarity); and, furthermore, evolutionary homologues do not necessarily even have a great deal of phenetic resemblance (van Valen 1982). The fundamental problem with automated alignment, then, is that the resulting hypotheses of evolutionary homology are frequently not plausible.

It may be possible to produce a plausible alignment by hand when there are apparently relatively few gaps needed to align the sequences (i.e. the alignment is apparently optimal under all conditions), such as in closely related sequences. Under these circumstances, there will be a high per-cent isology, and maximising sequence identity will be effective at detecting homology as well as straightforward to carry out. However, this is almost never true of non-translated sequences such as rDNAs or introns, where there are usually large differences in sequence length, especially as the sequences become less similar to each other. It is then necessary to introduce large numbers of gaps into the sequences to equalise their lengths, and the location of these gaps is not straightforward to determine. Note, incidentally, that it is not the unequal lengths that necessitates the gaps, but the low similarity—even equal-length sequences may need gaps in order to align homologues.

In situations where there is low per-cent isology between the sequences, it has been usual to use a mathematical algorithm to produce the alignment, because under these circumstances 'eyeball' alignment will be time-consuming, tedious, irreproducible and often biased (Henikoff 1991; Thorne *et al*. 1991). Interestingly, sequence divergence is good for tree building, since character variation is what provides support for the inferred branches on the tree (even homoplasy can provide branch support), but it is not good for alignment, since this is what makes the process difficult, and yet successful tree building is dependent on successful alignment.

So, phenetic pattern-matching algorithms are most commonly used, although other alignment strategies exist (as discussed in the next section). Clearly these algorithms, if based solely on the primary sequence information, can only succeed at uncovering evolutionary relationships to the extent that sequence similarity is the result of homology in the particular set of sequences being aligned. So, it is usually recognised that the 'correct' alignment may differ from the 'optimal' alignment. Unfortunately, even the mathematically optimal pairwise alignment will have statistical limitations on its accuracy (Holmes and Durbin 1998).

These pattern-matching algorithms all attempt to produce a sequence alignment that optimises some chosen criterion

of match between the individual sequences (an objective function or overall cost). That is, the sequences are compared by a pattern-matching process that searches for correspondence between the elements of the sequences, introducing gaps into the sequences as required to optimise some criterion for correspondence (usually minimising the cost, which is the sum of the weights applied to mismatches and gaps). From this point of view, the objective is to get the gaps in the right place—if each gap represents an indel (or sequencing error) then we have the evolutionarily correct alignment.

There are many algorithms currently available (reviewed by Gotoh 1999; Duret and Abdeddaïm 2000; Phillips *et al*. 2000; Nicholas *et al*. 2002; Notredame 2002; Lambert *et al*. 2003; Batzoglou 2005; Wallace *et al*. 2005*a*; Edgar and Batzoglou 2006; see Apostolico and Giancarlo 1998 for an overview of the earlier history). These optimise a variety of mathematical functions measuring the overall alignment cost, such as sum-of-pairs (sum of pairwise similarities), entropy (variation within an alignment column), log-expectation (a profile probability score) or consistency (agreement with a list of constraints). They operate in various modes, such as simultaneous, progressive, exact, stochastic (or non-deterministic, so that different runs of the program may produce different alignments) and iterative (so that the alignment is produced by a series of refinements). They are mostly based on dynamic programming (a process described by Wheeler 1994; Phillips *et al*. 2000; Phillips 2006) although other strategies exist. Unfortunately, the multiple-sequence alignment problem is mathematically NP-complete in all of its various guises (Wang and Jiang 1994; Wareham 1995; Bonizzoni and Della Vedova 2001; Just 2001; Elias 2003; Just and Della Vedova 2004; Kececioglu and Starrett 2004), and the number of possible alignments increases combinatorially with the length of the sequences and combinatorially again with the number of sequences (Slowinski 1998). So, there is no expectation that it will ever be practical to align many sequences simultaneously to find the globally optimal solution.

Therefore, for more than two sequences most of the alignment algorithms use exact procedures (which guarantee to find the optimal solution) to align the sequences pairwise, but then use heuristic procedures (computationally efficient strategies that should produce a solution that is at least close to the optimal one) to progressively braid these pairwise alignments into a multiple alignment (Hogeweg and Hesper 1984). These progressive-alignment procedures do not guarantee to produce the globally optimal alignment, because both the order in which the sequences are added and the optimisation path that the procedure chooses can affect the result (Wheeler 1994). Thus, there may be many equally optimal solutions, and misalignments made early in the process cannot subsequently be corrected (i.e. 'once a gap always a gap'; Feng and Doolittle 1987).

Variation in the outcome of the computerised sequence-alignment process thus occurs in at least two distinct ways. First, the different algorithms can produce different alignments. This is because they adopt different objective functions, and they have different heuristics for trying to optimise those functions. The phylogenetic trees resulting from these alignments can be on average more dissimilar to each other than are the trees produced by different tree-building methods (Morrison and Ellis 1997; Ogden and Rosenberg 2006). This is an important point, because most attention to date has been focused on variation caused by differences in the method of phylogenetic inference (tree building, which assesses secondary homology) rather than in the assessment of topographic identity (sequence alignment, which assesses primary homology).

Second, most alignment algorithms have a series of available parameters that can be varied. For example, protein alignments require decisions concerning the relative costs of the different substitutions among the various amino acids, and nucleotide alignments can require relative costs for transitions and transversions. However, probably the most important of the parameters are the alignment-gap cost-ratios (gap weights or penalties), which refer to the relative cost of inserting a new gap into a sequence or extending an already-existing gap compared to a substitution. Ideally, the relative costs should reflect the probability of the indel events relative to substitution events (Wheeler 1993), but in practice there is no objective criterion for choosing the costs (Vingron and Waterman 1994), there is no way of determining analytically what these costs should be (Rinsma-Melchert 1993), and in spite of much analysis there is little empirical guidance (Britten *et al*. 2003; Kececioglu and Kim 2006).

Thus, the computer programs that implement the alignment algorithms usually have default values for the costs that are designed to produce 'biologically interesting' results, in the sense that an effort has been made to use biological data to infer meaningful values (e.g. based on typical globular proteins). There is, however, no reason to assume that the authors of the programs have optimised the choice of these values for any particular purpose (Blackshields *et al*. 2006). Very few biologists seem to be willing to deviate from these default choices, but the assessments published to date show that the values chosen can have significant effects for the alignment of amino acids (Fitch and Smith 1983; Henneke 1989; Tyson 1992; Taylor 1996) and nucleotides (Fitch and Smith 1983; Wheeler 1995; Milinkovitch *et al*. 1996; Titus and Frost 1996; Morrison and Ellis 1997; Cerchio and Tucker 1998; O'Brien *et al*. 1998; Smith and Hurst 1998; Sanchis *et al.* 2001; Terry and Whiting 2005), and that these effects increase as the sequences become less similar to each other. Furthermore, gap costs are likely to differ between different gene types; for example, in a protein-coding gene even a small gap may create a frame-shift that has a large effect on the function of the translated product, while for an RNA-coding

gene the same gap can, at worst, affect a single functional motif. It has also been argued that even for a single gene there can be no fixed costs, because the probability of indels varies from one region of the sequence to another (Kjer 1995, 2004). Several programs exist to explore the sensitivity of an alignment to variation in these parameter values (Yuan *et al*. 1999; Löytynoja and Milinkovitch 2001), and this approach (called sensitivity analysis) could be more usefully employed to assess uncertainty arising from the optimisation parameters (Giribet and Wheeler 1999; Phillips *et al*. 2000; Terry and Whiting 2005), although this should not be mistaken for an assessment of uncertainty (or robustness) in the alignment itself (Redelings and Suchard 2005). Clearly, it is inadequate to simply report that a particular computer program was used to align the sequences, without also reporting the parameter values used.

There is now much empirical evidence that multiple-sequence alignments produced by the pattern-matching programs are not necessarily similar to those based on structure or function considerations, both for amino acids (Taylor 1986; Barton and Sternberg 1987; Johnson *et al*. 1990; Gotoh 1996; Briffeuil *et al*. 1998; Thompson *et al*. 1999*b*; Katoh *et al*. 2002, 2005*a*, 2005*b*; Lassmann and Sonnhammer 2002; Marchler-Bauer *et al*. 2002; Raghava *et al*. 2003; Edgar 2004*b*; Van Walle *et al*. 2004; Yamada *et al*. 2004; Do *et al*. 2005; Subramanian *et al*. 2005; Simossis *et al*. 2005; Zhou and Zhou 2005; Sze *et al*. 2006) and for nucleotides (Ellis and Morrison 1995; Kjer 1995; Morrison and Ellis 1997; Beebe *et al*. 2000; Hickson *et al*. 2000; Mugridge *et al*. 2000; Gardner *et al*. 2005; Katoh *et al*. 2005*b*; Lebrun *et al*. 2006). It is therefore worthwhile to explore this problem in more detail. Most of the detailed information will relate to amino acid sequences, since this is where it has accumulated, but the same general principles will apply to nucleotide sequences as well.

Simple sequence alignment

Far and away the most commonly used progressive-alignment computer program in molecular phylogenetic studies is Clustal, either in the character-based version W (Thompson *et al*. 1994) or the graphics-based version X (Thompson *et al*. 1997). This is a justified choice for phenetic alignment, as this program uses a relatively simple and quick algorithm that can be very effective, making it the benchmark standard for such programs for nearly 20 years (Chenna *et al*. 2003). For this reason, I will use this program as the basis for a discussion of the limitations of pattern-matching programs for phylogenetic sequence alignment. It is important that we understand these limitations because they will have a serious effect on our ability to obtain a multiple alignment that is usable for phylogenetic purposes.

Clustal uses a heuristic device to approximate the global similarity alignment. It employs a sum-of-pairs criterion to evaluate the multiple alignment, the objective being to maximise the sum of the pairwise similarities between the aligned sequences, based on a cost matrix that quantifies the similarity of every pair of residues. Since finding the global optimum is an NP-hard problem, approximations are used to find a solution that is near to the optimum without guaranteeing to have found it, the basic heuristic device being progressive rather than simultaneous evaluation of the pairwise alignments. Thus, there is an optimality criterion for the procedure, but there is no way of knowing how close any result is to the optimum result.

The program first performs all possible pairwise alignments using the dynamic programming algorithm. From these alignments a matrix of all possible pairwise distances is calculated, and a neighbour-joining tree is produced. This acts as a guide tree to determine the order in which the sequences enter the multiple-alignment process, working from the tips towards the root. The multiple alignment is produced by dynamic programming using a combination of sequence–sequence alignment, sequence–profile alignment and profile–profile alignment. The substitution cost matrix used varies depending on the degree of sequence divergences, the gap penalties depend on how many other sequences have a gap at the same position and on the solvent accessibility of the nearby amino acids (if relevant), and the sequences are weighted inversely to how many close relatives they have in the alignment. This overall algorithm produces a flexible alignment procedure that is very efficient in straightforward situations, especially for amino acid alignments, which are treated in a somewhat more sophisticated manner than are nucleotide alignments.

Note that most of this sophistication comes from trying to put some biological insight into the procedure, to counter-balance the simple idea of maximising sequence similarity. For example, the use of substitution-cost matrices for amino acids (such as the BLOSUM, Gonnet and PAM series) is an attempt to quantify evolutionary relatedness (Vogt *et al*. 1995). The estimates of exchangeability among the residues contained in these matrices is an attempt to quantify the likelihood of evolutionary relationships among the amino acids (they could, for example, be based on physico-chemical properties, instead). This is actually a brave attempt to treat mathematically the nebulous concept of substitutions as unobservable historical events; we reach the limits of mathematics when we want to reconstruct unique historical events, and we end up with probabilities instead. Little has been done along this line for nucleotide sequences, although such matrices could be based on variability due to the transition : transversion ratio and nucleotide frequencies (Chiaromonte *et al*. 2002). Moreover, the use of gap penalties that vary depending on the adjacent amino acids is an attempt to avoid placing gaps in secondary structure features that are known to be relatively free of gaps, such as helices and sheets or solvent-inaccessible regions of proteins (Henneke 1989). Thus, the program cannot be accused of being free of

biological insight, even if that insight is always implemented through the idea of similarity.

Unfortunately, in many complex situations progressive-alignment programs have serious limitations, both for amino acid and nucleotide alignment. Complexity is largely created by sequence divergence—the less similar the sequences are to each other then the harder it is to align them, since the alignment procedure is based solely on similarity. There are many patterns for divergence among sequences, but we can consider the simplest one first. This is where the sequences are all approximately equally similar (or dissimilar) to each other. Under these circumstances, decreasing similarity creates decreasing accuracy of the multiple alignment (Simmons and Freudenstein 2003). An example of this is shown in Fig. 4, where decreasing identity among the sequences measures divergence. (Note that, strictly speaking, sequence identity can only be measured on the true alignment; Gardner *et al*. 2005; Rosenberg 2005*a*. It also requires specification of the denominator in the calculation, which could be, for example, length of the alignment, number of non-gap positions, length of the shortest sequence, or mean length of the sequences; May 2004. I have used the number of non-gap positions.)

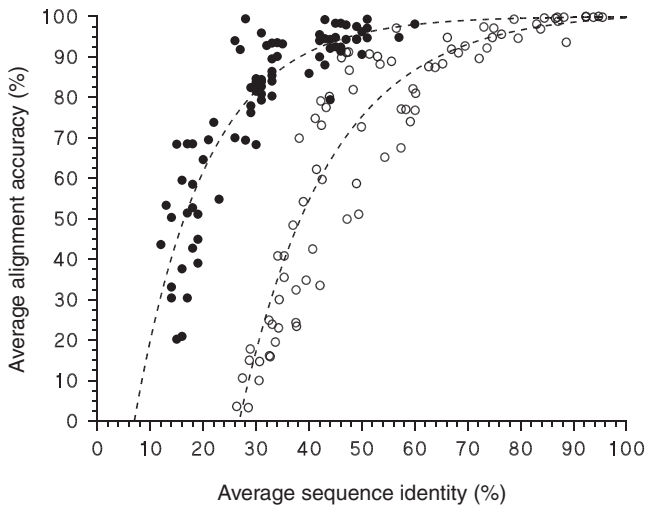

Fig. 4. Two examples of the relationship between sequence identity and alignment correctness, showing the dramatic decrease in accuracy as sequence identity decreases. Each solid point represents a single amino-acid alignment from Reference Set 1 of the BAliBASE set of manually curated alignments (Thompson *et al.* 1999*a*); these reference alignments are assumed to be correct based on 3D structural superpositions of the proteins. Each open point represents a single nucleotide alignment generated by the Rose simulation program (Stoye *et al.* 1998), which simulates the evolution of sequences under a simple artificial model; the HKY substitution model was used to create eight sequences each 250 bases long, with nucleotide frequencies set to those of coding sequences in the current public databases, and indel frequencies and sizes based on those of Gu and Li (1995). The multiple alignments for comparison in both cases were performed with the ClustalW 1.83 program, with default settings; and the score shows the average percentage of alignment positions that were correctly aligned by comparison with the reference alignment.

The consensus opinion from those studies done to date on the accuracy of progressive-alignment programs such as Clustal seems to be something like Table 1 for amino acid sequences. The region of 20–40% identity is known as the Twilight Zone (Doolittle 1981), where homologous pairs of proteins usually show structural similarity but not primary sequence similarity. Here, alignment accuracy can be unpredictably anywhere between 0 and 100% if based on similarity alone (Fig. 4). The region of <20% identity is the Midnight Zone, where sequence alignment is likely to be no better than random. In fact, if all amino acids are equally abundant in two sequences then perfectly random alignment will occur at 5% identity; for the relative amino acid abundances in the current public databases the expected value is 6–7% identity. Not surprisingly, the empirical Dayhoff model of protein evolution (on which the well known PAM matrices of amino acid similarity are based) reaches equilibrium at this same point (Higgins *et al*. 1996). Successful structure-based pairwise alignments are known down to ~3% amino acid identity, however (Stebbings and Mizuguchi 2004).

Clearly, only those Clustal alignments at >80% amino acid identity are likely to be acceptable for phylogenetic analysis without some sort of intervention to supplement the sequence-similarity information. (Note that Clustal version 1.83 can output a percent identity matrix, so that you can check whether this situation applies to your analysis; Chenna *et al*. 2003.) Alternative alignment strategies should then be preferred. As discussed in the next section, alignment methods appear to have the following order of accuracy: profile–profile > hidden markov models > sequence–profile > dynamic programming > heuristic (Marti-Renom *et al.* 2004).

Similar qualities apply to nucleotide sequences, although the situation is quantitatively different. For example, randomly aligned nucleotide sequences will have ≥25% identity, depending on the base composition. Given the relative nucleotide frequencies in the current public databases (G ≈ C ≈ 21%, A ≈ T ≈ 29% for coding sequences), an alignment between two random sequences will have ~27% identity. Thus, the form of the curve for nucleotides shown in Fig. 4 is similar to that for amino acids, but is shifted to the right to a location that depends on the

Table 1. Accuracy of progressive alignment programs for amino acid sequences

Amino acid identity (%)	Alignment accuracy (%)
>80	>95
>60	>90
>40	>80
<20	<80

nucleotide composition of the sequences being aligned, with the Twilight Zone at 40–50% (cf. Gardner *et al*. 2005). Thus, the accuracy of progressive-alignment programs such as Clustal seems to be something like Table 2 for nucleotide sequences. As you can see, it is generally more difficult to align nucleotide sequences than amino acid sequences—as a generalisation, alignment difficulty at 50% identity for amino acids might correspond to 70% for nucleotides (Duret and Abdeddaïm 2000).

The basic limitation for trying to assess sequence homology is that the current tree-building analyses are based on the primary sequence data only. So, low percent identity among sequences means that tree building can fail even if the homology assessment is perfectly correct (i.e. the hypotheses of homology cannot be assessed on the tree effectively). There will thus be a practical lower limit for per-cent identity below which a sequence alignment is useless for phylogenetic purposes, as the data patterns will be effectively random. Sequence alignment in the Twilight Zone thus may be useful only for database searching, sequence comparison and structure prediction. Under these circumstances a phylogenetic analyses needs to be based on other data types, such as gene order or distances between secondary structures.

What I have said so far applies only to the simplest situation for patterns of divergence among sequences (i.e. approximately equal similarity among the sequences). Many other possible patterns are known to affect the success of sequence alignment even more than simple decreasing similarity. These include sets of sequences where there are: (i) a small number of 'orphan' sequences, which do not have close similarity to the remaining sequences (Fig. 5*c*); (ii) a series of distinct sequence subgroups that have high similarity within subgroups but not between subgroups (Fig. 5*d*); (iii) long-terminal or internal gaps (Fig. 5*a*, *b*); and (iv) sequence complexities such as repeats, translocations and inversions (Table 3). In order of difficulty for progressive-alignment programs these are likely to be (iv) > (ii) > (iii) > (i). Note that situations (i) and (ii) will correspond to the situation where the outgroup taxa are quite different from the ingroup taxa, and so they might be common in phylogenetic studies. Indeed, they probably constitute a major part of the reason why so many problems have been identified with the use of distant outgroups in tree-building analyses (Morrison 2006). Situation (iii) is often a by-product of using different primers for sequencing different taxa, and so is also quite common in phylogenetic studies. Sequence regions subject to (iv) are usually actively avoided in phylogenetic studies, at least partly because of the extreme difficulty of analysing the data (see Pei and Grishin 2006).

Table 2. Accuracy of progressive alignment programs for nucleotide sequences

Nucleotide identity (%)	Alignment accuracy (%)
>90	>95
>80	>90
>70	>80
<60	<80

Figure 6 shows a specific example of the sorts of problems that can arise in practice when using a progressive alignment program. The sequences are taken from the seed amino-acid alignment of the *gp120* family in the Pfam protein-domain database. *Gp120* is the crucial envelope protein of HIV (lentiviruses, a subfamily of the retroviruses) that facilitates binding to and fusion with the target cells, which are human CD4 lymphocytes. The figure compares a progressive alignment with the structure-based alignment, as I have done throughout this section. The basic conclusion is that the hypotheses of evolutionary events are far more parsimonious and plausible for the structure-based alignment than for the similarity-based alignment.

In this example, the Clustal program does quite well with most of the conserved parts of the sequences, but not with the five variable regions. The success occurs because the program has various weighting factors inbuilt, which allow it to detect structural features such as helices, sheets and bridges, and to preferentially avoid putting gaps into these locations. However, this procedure is not always successful. There are two alignment blocks illustrated in Fig. 6, showing that most of the alignment is structurally correct but with several notable errors. For example, the first disulfide bridge (positions 6–7) is not conserved in this alignment (although the other two are), as the HIV-1 and HIV-2 sequences are misaligned against each other. This mistake occurs at the final step of the profile-alignment iterations, where the two main sequence groups are finally aligned against each other; and failure to align such conserved residues is considered to be the most common error made by progressive-alignment programs (Nicholas *et al*. 2002). Second, the second helix (positions 10–30) has a two-amino-acid gap inserted into it near the end, causing misalignment. Note that the structure alignment has a one-amino-acid gap near the beginning of the helix, indicating that the HIV-1 and HIV-2 proteins do not have identical structures. Third, for the first β-sheet (positions 40–45) the HIV-1 and HIV-2 sequences are correctly aligned within each taxonomic group but not between groups. Fourth, the final helix (the second alignment block) starts correctly but becomes misaligned. This is because the HIV-1 and HIV-2 proteins do not have the same structure at this point, requiring a gap to be inserted into the helix. However, ClustalW preferentially down-weights terminal gaps, so that it rarely inserts internal gaps near the ends of an alignment. This is a known limitation of the program, and it has the same problem at the beginning of its alignments as well as at the end.

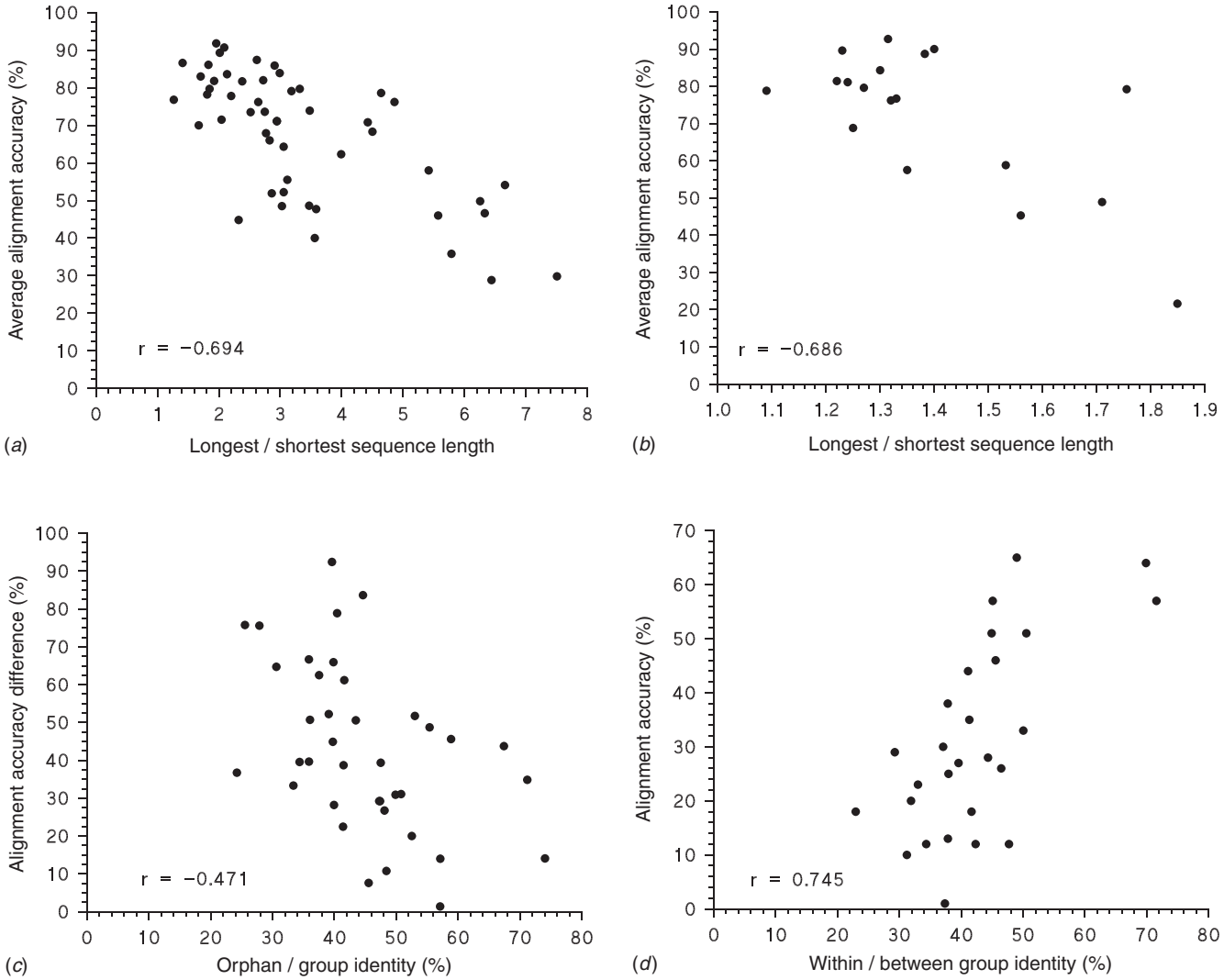

Fig. 5. Four examples of the effect of sequence characteristics on the success of progressive multiple sequence alignment, illustrating the circumstances under which similarity-based alignments will fail. Each point represents a single amino-acid alignment from one of the pooled Reference Sets of the structure-based BAliBASE versions 2 and 3 manually curated alignments (Bahr *et al.* 2001; Thompson *et al.* 2005); these reference alignments are assumed to be correct based on 3D structural super-positions of the proteins. The multiple alignments for comparison were performed with the ClustalW 1.83 program, with default settings. (*a*) Average percentage of pairwise alignment positions that were correctly aligned for sequences with *N*/*C*-terminal extensions (or deletions); the length ratio is shown only for the range 1–8 (the original data range to 35). (*b*) Average percentage of pairwise alignment positions that were correctly aligned for sequences with internal insertions or deletions; the length ratio is shown only for the range 1–2 (the original data range to 6). (*c*) Decrease in percentage of alignment positions that were correctly aligned for sequences with 1–4 highly divergent 'orphan' sequences; the abscissa is the ratio of average % identity between the orphan sequences and the other sequences to average percentage identity among the other sequences. (*d*) Percentage of alignment positions that were correctly aligned for sequences containing 2–5 subgroups with <25% residue identity between groups; the abscissa is the ratio of average percentage identity within the subgroups to average percentage identity between the subgroups.

Gap costs and character definition

From a theoretical perspective, the main reason for the failure of similarity-based alignment is that it does not provide all three of the characteristics that science requires: description, prediction, and explanation. The alignment procedure tries solely to provide an efficient description of the patterns in the data, which it does by providing a parsimonious alignment of the residues. This alignment can then be used by a scientist as an implicit set of predictions about homology of the residues, but the alignment procedure itself does not explicitly try to provide this (i.e. it describes patterns without regard to how they were formed) and therefore it comes as no surprise that it frequently fails. Moreover, the models used by the algorithm do not try to model sequence evolution at all (i.e. they model sequence patterns rather than historical processes), and so the resulting alignments do not necessarily provide any

Table 3. Accuracy of the ClustalW program for BAliBASE reference sets 6–8

The data refer to the decrease in average percentage of pairwise alignment positions that were correctly aligned by comparison with the structure-based BAliBASE reference alignment (Bahr *et al*. 2001). The repeats have been classified into several subtypes according to their residue similarity. For each dataset, the repeated sequence was aligned alone, to provide the reference degree of alignment accuracy, and then the complete sequences were aligned. The reduction in accuracy is the difference in accuracy between the whole alignment and the repeat alignment as a percentage of the accuracy of the repeat alignment. Only datasets with average sequence identity >80%, and the repeat sequence length <50% of the whole sequence, were included

Description of sequence pattern	Sample size	Reduction in accuracy (%)
The same number of repeats of a unique subtype	4	29.6
A variable number of repeats of a unique subtype	5	34.5
The same number of repeats with different subtypes in the same order	11	35.0
The same number of repeats with different subtypes, but in a different order	17	50.7
A variable number of different repeat subtypes	18	36.7
The presence of an additional non-repeated conserved domain	8	30.5
The presence of more than one different repeat type	8	79.3
Inverted domains	4	62.2

```
Similarity alignment
            1       10        20        30        40        50        60        70        1       10
ENV_HV1A2   IRKAHCNISRAQWNNTLEQIVKKLR--EQFG-----NNKTIVFNQSSGGDPEIVMHSFNCRGEFFYCNTTQLFN   KAKRRVVQREKR----
ENV_HV1B1   MRQAHCNISRAKWNNTLKQIDSKLR--EQFG-----NNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFN   KAKRRVVQREKR----
ENV_HV1C4   IRQAHCNISRAQWNNTLQQIATTLR--EQF------GNKTIAFNQSSGGDPEIVMHSFNCGGEFFYCNSTQLFN   KAKRRVVQREKR----
ENV_HV1EL   IGQAHCNISRAQWSKTLQQVARKLG--TLL------NKTIIKFKPSSGGDPEITTHSFNCGGEFFYCNTSGLFN   RAKRRVVEREKR----
ENV_HV1RH   IRKAHCNLSRAQWNNTLKQVVTKLR--EQF------DNKTIVFTSSSGGDPEIVLHSFNCGGEFFYCNTTQLFN   RAKRRVVQREKR----
ENV_HV1W1   IRQAHCNISRAKWNNTLKQIVEKLR--EQF------KNKTIVFNHSSGGDPEIVTHSFNCGGEFFYCDSTQLFN   KAKRRVVQREKR----
ENV_HV1Z8   IRQAYCNISAAAWNKTLQQVAKKLG--DLL------NQTTIIFKPPAGGDPEITTHSFNCGGEFFYCNTSRLFN   RAKRRVVEREKR----
ENV_SIVCZ   TRSAYCKINGTTWNRTVEEVKKALA--TSSNR----TAANITLNRASGGDPEVTHHMFNCGGEFFYCNTSQIFT   KARRHTVARQKDRQKR
ENV_HV2BE   RPRQAWCRFGGRWREAMQEVKQTLVQHPRYKG-INDTGKINFTKPGAGSDPEVAFMWTNCRGEFLYCNMTWFLN   DQRR-SSTPV-RNKR-
ENV_HV2CA   RPRQAWCWFKGNWTEAMQEVKQTLAEHPRYKG-TKNITDITFKAPERGSDPEVTYMWSNCRGEFFYCNMTWFLN   SQKRYSPAHG-RPKR-
ENV_HV2D1   KPGQAWCWFQGNWIEAMREVKQTLAKHPRYGG-TNDTGKINFTKPGIGSDPEVTYMWTNCRGEFLYCNMTWFLN   KEKRYSSAPV-RNKR-
ENV_HV2G1   RPRQAWCWFKGKWREAMQEVKQTLIKHPRYKG-TNDTKNINFTKPGRGSDPEVAYMWTNCRGEFLYCNMTWFLN   REKRYSSAPV-RNKR-
ENV_HV2NZ   KPRQAWCWFEGQWKEAMQEVKETLAKHPRYKGNRSRTENIKFKAPGRGSDPEVTYMWTNCRGESLYCNMTWFLN   SVKRYSSAHQ-RHTR-
ENV_SIVM1   RPKQAWCRFGGNWKEAIKEVKQTIVKHPRYTG-TNNTDKINLTAPR-GGDPEVTFMWTNCRGEFLYCKMNWFLN   NVKRYTTGGTSRNKR-

Structure alignment
Structure  .HHH..D...HHHHHHHHHHHHHHHHHHH..........SSSSSS.......SSSSSSS.D..SSSSD.......   .HHHHHHH...HHHH...
ENV_HV1A2   IRKAHCNISRAQWNNTLEQIVKKLREQFGNN------KTIVFNQSS-GGDPEIVMHSFNCRGEFFYCNTTQLFN   KAKRRVVQ---REKR
ENV_HV1B1   MRQAHCNISRAKWNNTLKQIDSKLREQFGNN------KTIIFKQSS-GGDPEIVTHSFNCGGEFFYCNSTQLFN   KAKRRVVQ---REKR
ENV_HV1C4   IRQAHCNISRAQWNNTLQQIATTLREQFG-N------KTIAFNQSS-GGDPEIVMHSFNCGGEFFYCNSTQLFN   KAKRRVVQ---REKR
ENV_HV1EL   IGQAHCNISRAQWSKTLQQVARKLGTLLN-K------TIIKFKPSS-GGDPEITTHSFNCGGEFFYCNTSGLFN   RAKRRVVE---REKR
ENV_HV1RH   IRKAHCNLSRAQWNNTLKQVVTKLREQFD-N------KTIVFTSSS-GGDPEIVLHSFNCGGEFFYCNTTQLFN   RAKRRVVQ---REKR
ENV_HV1W1   IRQAHCNISRAKWNNTLKQIVEKLREQFK-N------KTIVFNHSS-GGDPEIVTHSFNCGGEFFYCDSTQLFN   KAKRRVVQ---REKR
ENV_HV1Z8   IRQAYCNISAAAWNKTLQQVAKKLGDLLN-Q------TTIIFKPPA-GGDPEITTHSFNCGGEFFYCNTSRLFN   RAKRRVVE---REKR
ENV_SIVCZ   TRSAYCKINGTTWNRTVEEVKKALATSSNRTA-----ANITLNRAS-GGDPEVTHHMFNCGGEFFYCNTSQIFT   KARRHTVA---RQKDRQKR
ENV_HV2BE  RPRQAWCRFGG-RWREAMQEVKQTLVQHPRYKG-INDTGKINFTKPGAGSDPEVAFMWTNCRGEFLYCNMTWFLN   DQRRYSSTPV-RNKR
ENV_HV2CA  RPRQAWCWFKG-NWTEAMQEVKQTLAEHPRYKG-TKNITDITFKAPERGSDPEVTYMWSNCRGEFFYCNMTWFLN   SQKRYSPAHG-RPKR
ENV_HV2D1  KPGQAWCWFQG-NWIEAMREVKQTLAKHPRYGG-TNDTGKINFTKPGIGSDPEVTYMWTNCRGEFLYCNMTWFLN   KEKRYSSAPV-RNKR
ENV_HV2G1  RPRQAWCWFKG-KWREAMQEVKQTLIKHPRYKG-TNDTKNINFTKPGRGSDPEVAYMWTNCRGEFLYCNMTWFLN   REKRYSSAPV-RNKR
ENV_HV2NZ  KPRQAWCWFEG-QWKEAMQEVKETLAKHPRYKGNRSRTENIKFKAPGRGSDPEVTYMWTNCRGESLYCNMTWFLN   SVKRYSSAHQ-RHTR
ENV_SIVM1  RPKQAWCRFGG-NWKEAIKEVKQTIVKHPRYTG-TNNTDKINLTAPR-GGDPEVTFMWTNCRGEFLYCKMNWFLN   NVKRYTTGGTSRNKR
```

Fig. 6. Two parts of the seed amino-acid alignment of the *gp120* family from the Pfam protein-domain database (Finn *et al.* 2006), each aligned in two different ways, demonstrating the problems that can arise in progressive alignments. The sequences are from selected strains of Human Immunodeficiency Virus One (HIV-1), Human Immunodeficiency Virus Two (HIV-2) and Simian Immunodeficiency Virus (SIV). The first alignment shows two blocks produced by ClustalW 1.83, using default parameters. The second alignment is from the Pfam database, also showing the protein structure associated with the sequences. There are 10 α-helices in the sequences (two shown in the figure as H), 30 β-sheets (three shown as S) and 9 disulfide bridges (part of three pairs shown as D). There are also five variable regions (loops V1–V5), which are not shown.

explanatory insights into evolutionary processes. If science requires accurate descriptions, explicit predictions and plausible explanations, then similarity-based alignment can potentially fall down on both the second and third criteria.

One of the main practical areas where this failure becomes evident is the treatment of gap costs. As noted above, these are usually treated as being of the form: $\text{cost} = a + b \times \text{length}$ (i.e. a gap-opening cost plus a length-dependent gap-extension cost), which is known as an affine gap cost (Gotoh 1982). The problem is that this approach does not model the evolutionary processes that create gaps in an alignment (i.e. indels). I have already emphasised that the alignment programs recognise residues (nucleotides or amino acids) as the 'characters' in sequence data. It is straightforward to model substitution events when

residues are the characters, because these events affect only one character at a time. However, it is hard to treat indels as characters from this perspective, because when an indel event occurs it may simultaneously affect more than one nucleotide position in a sequence. Indel characters and substitution characters are actually quite different things, and it is hard to create a model that incorporates them both. So, the problem with indels in a model is that multiple-base indels span more than one substitutional character, and an indel model cannot be a simple extension of our current models. Furthermore, insertions and deletions may result from two different sorts of events (Chang and Benner 2004), with slipped-strand mispairing causing short indels and unequal crossing-over, transposition and inversion causing longer ones (>30 bases); and so an indel model may not be a straightforward concept.

A similar problem arises for all of the other mutational events that affect blocks of sequence simultaneously (i.e. subsequences), such as inversions, translocations, transpositions and tandem duplications. The treatment of each alignment position as a single character does not treat each of these as a single evolutionary event, but instead treats each positional difference as a different substitution. That is, the characters are treated as being independent of each other when they are not. This creates a heavy weighting against alignment of the subsequences involved, which makes it difficult for the alignment programs to correctly detect and align these events even in a pairwise alignment (Benson 1997; Sammeth *et al*. 2005), let alone a multiple alignment (Raphael *et al*. 2004; Wegner *et al*. 2004; Sammeth and Heringa 2006).

Instead of modelling indels as a separate concept, the use of an affine gap cost actually models the gaps as two different types of substitutions, one with a high cost (for the gap-opening residue) and one with a low cost (for the gap-extension residues). The end result is that the pattern-matching programs add fewer and/or shorter gaps to the alignment than the number of indels expected, especially for distantly related sequences (Thorne and Kishino 1992; Morrison and Ellis 1997; Nicholas *et al*. 2002; Simmons and Freudenstein 2003; Löytynoja and Goldman 2005). This problem is compounded for methods based on sum-of-pairs, as these weight some evolutionary events more strongly than others (Gonnet *et al*. 2000); and the failure of affine gap costs to deal with long gaps also violates the triangle inequality, which leads to trivial alignments (Aagesen *et al*. 2005). There have been several empirical studies indicating the extent to which affine gap costs underestimate the probability of long indels in different types of sequences (Pascarella and Argos 1992; Benner *et al*. 1993; Gu and Li 1995; Ophir and Graur 1997; Graham *et al*. 2000; Qian and Goldstein 2001; Reese and Pearson 2002; Chang and Benner 2004; Keightley and Johnson 2004; Wrabl and Grishin 2004); that is, the frequency distribution of gaps in real sequences has a much longer right-hand tail than the geometric distribution modelled by the affine gap cost. This situation becomes progressively worse as the sequence identity decreases, which is what leads to the decreased performance of the similarity-based programs. Non-linear gap costs have not been used, even though they would obviously be more appropriate (e.g. the more biologically reasonable cost $= a + b \times \log[\text{length}]$), because this would increase the time requirement of the pairwise dynamic programming algorithm from the square of the sequence length to the cube (Fitch and Smith 1983). It might thus be better to model the number of gaps rather than the gap lengths (Nozaki and Bellgard 2005).

Part of the problem here is the methodological convention of parsimony. The similarity programs are based on the idea of descriptive parsimony: inserting the minimum number and length of gaps that are consistent with the substitution model being used. However, this convention obscures the fact that the descriptive parsimony is actually based on ontological parsimony (Johnson 1982). That is, the justification for employing descriptive parsimony is to assume that evolution itself has been parsimonious (i.e. that evolutionary processes are parsimonious in their use of gaps). However, there is no empirical evidence that evolution acts parsimoniously and, indeed, quite a lot of evidence to the contrary. Therefore, use of descriptive parsimony only ensures that the 'true' situation will not be simpler than we have hypothesised; that is, there will not be fewer or shorter gaps than the similarity-based programs produce. We should not be surprised, then, that gaps are longer in reality than the programs suggest. This does mean, unfortunately, that the resulting alignments are wrong, even if we do know that we have underestimated the amount of evolution that has occurred.

In this regard, it is interesting to consider whether finding the globally optimal similarity alignment is worthwhile in the first place. Programs such as Clustal do not guarantee to have found the optimum alignment, since this is not practical for the length and number of sequences that most biologists are dealing with. However, this may *not* be a negative feature if the globally optimal alignment is not the true alignment. To assess this, I have compared the ClustalW 1.83 alignments for the 82 BAliBASE version 2 Reference Set 1 sequences with those produced by either the OMA version 0.98 (Reinert *et al*. 2000) or the MSA version 2.1 (Gupta *et al*. 1995) programs. Although neither of these latter two programs can guarantee to find the globally optimal similarity alignment based on sum-of-pairs (the same optimality criterion that Clustal uses), they each make a serious attempt to do so, and I have accepted the optimal alignment as being the highest-scoring one produced by either of these programs. The results of the two programs were often identical (see also Reinert *et al*. 2000), but MSA did better than OMA for 11 of the 28 long sequences. The results of the comparison with Clustal (Fig. 7) show that Clustal generally produces alignments that are as close to the reference alignment as

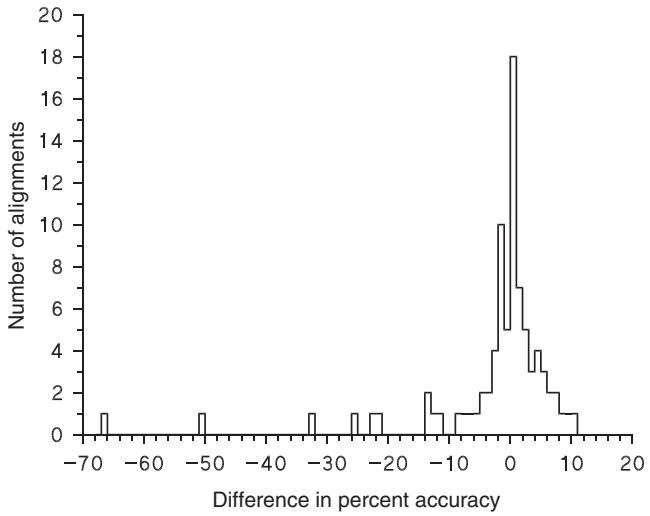

Fig. 7. Comparison of the accuracy of the globally optimal alignment (produced by either the MSA or OMA program with default settings) with that of a progressive alignment (produced by ClustalW 1.83 with default settings), with reference to a structure-based alignment. Each observation represents a single amino-acid alignment from reference set 1 of the BAliBASE set of manually curated alignments (Thompson *et al.* 1999*a*); these reference alignments are assumed to be correct based on 3D structural super-positions of the proteins. The abscissa shows the difference (ClustalW minus either OMA or MSA) in the average percentage of alignment positions that were correctly aligned by comparison with the reference alignment. The negative scores show those alignments for which the progressive alignment performed better than did the globally optimal alignment.

those from the other two programs, but that it can often do much, much better. (Note that Althaus *et al*. [2002] report similar results for their optimality program, COSA.) In other words, (i) Clustal's heuristic strategy mostly does a good job of finding the optimal sum-of-pairs alignment, and (ii) the suboptimal alignments from Clustal are as good as or better than the optimal alignments, because the reference alignments are not optimal in terms of similarity, but are based on structural considerations. A similar idea applies to phylogenetic alignment, because (as pointed out by Phillips 2006) a globally optimal alignment is effectively based on an unresolved 'star' tree, and thus contains no explicit evolutionary information. Suboptimal similarity may thus be closer to reality than is 'optimality'.

Improving automated alignment

Assessment of computer programs

The limitations of simple progressive sequence alignment have long been known (although BAliBASE seems not to have been used to illustrate them, as I have done here). Thus, various improvements to the simple progressive alignment algorithm have been developed, designed to deal with the fundamental limitation, which is that misalignments made early in the progressive process are not subsequently corrected. Several the associated computer programs have been compared for their ability to detect specified conserved sequence motifs (McClure *et al*. 1994; Briffeuil *et al.* 1998; Hudak and McClure 1999; Hickson *et al*. 2000) and for the overall match of their alignments to structure-based alignments (Gotoh 1996; Morrison and Ellis 1997; Thompson *et al*. 1999*b*; Notredame *et al*. 2000; Karplus and Hu 2001; Katoh *et al*. 2002, 2005*a*, 2005*b*; Lassmann and Sonnhammer 2002; Raghava *et al*. 2003; Edgar 2004*b*; Grasso and Lee 2004; Van Walle *et al*. 2004; Do *et al*. 2005; Subramanian *et al*. 2005; Simossis *et al*. 2005; Zhou and Zhou 2005; Sze *et al*. 2006) or to simulated alignments (Pollard *et al*. 2004; Ogden and Rosenberg 2006). None of the programs are completely successful based on any of the three criteria, and their success rate can sometimes be very low.

Rather than provide a new assessment of these alternative multiple-alignment programs, I will simply summarise some of the previously published results, and then introduce a few of the programs that seem to have direct relevance for phylogenetic analyses. Fig. 8*a* summarises the results for the PREFAB amino-acid alignment database, showing that several of the programs are notably less affected than Clustal by decreasing sequence identity, particularly in the Twilight and Midnight zones of amino-acid alignment. However, in practical terms there is little to choose between the improved programs.

Table 4 summarises the results for the BAliBASE and IRMBase amino-acid alignment databases, which relate to some of the other features that I identified in the previous section as affecting multiple alignments. [Katoh *et al*. (2005*b*), Blackshields *et al*. (2006), Roshan and Livesay (2006) and Wallace *et al*. (2006) provide recent comparisons of some later versions of these programs using BAliBASE version 3.] The BAliBASE database assesses global-alignment success while IRMBase assesses local-alignment success. Hence, only the two programs that incorporate local-alignment information (DiAlign and T-Coffee) perform well on the IRMBase alignments. For BAliBASE, the results are quite erratic, with most of the programs performing well on at least one Reference Set and badly on at least one other. The major stumbling blocks for most of the programs are Reference Set 4, consisting of sequences with N / C-terminal extensions, and particularly Reference Set 3, consisting of subgroups with <25% residue identity between groups; and so you should be particularly wary of these two characteristics if they appear in your datasets. Once again, there is little to choose between most of the improved programs.

The only recent evaluations of multiple-alignment programs with respect to nucleotide alignments are those of Gardner *et al*. (2005) and Katoh *et al*. (2005*b*) on the BRAliBASE database of structural RNAs. They conclude that success at amino-acid alignment is not necessarily a reliable guide to success at nucleotide alignment, at least

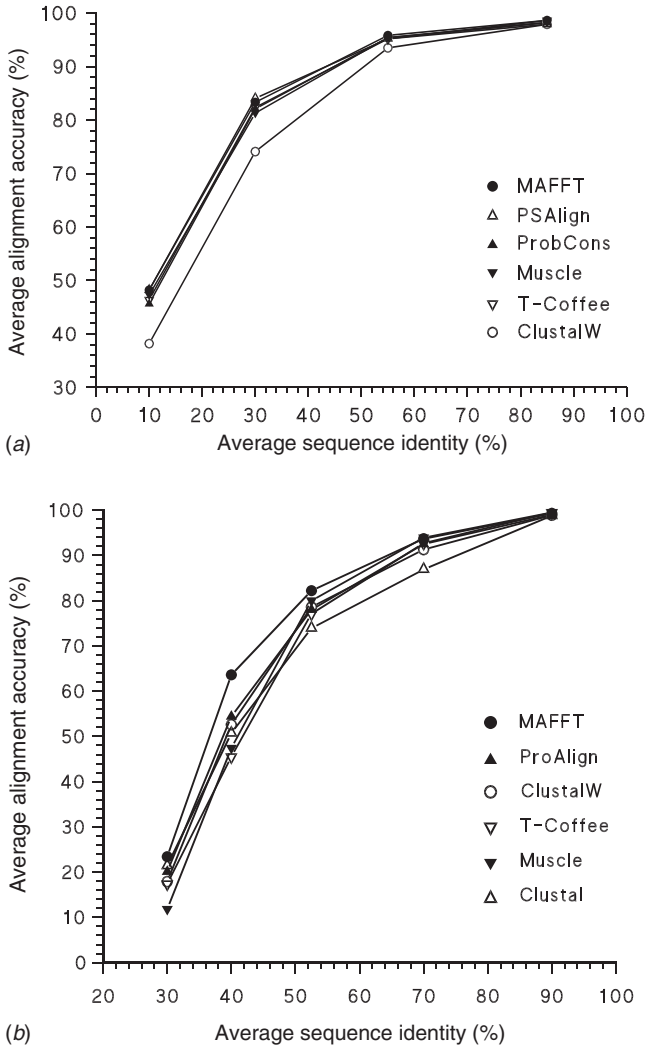

Fig. 8. Relationship between sequence identity and alignment correctness for several multiple-alignment programs, illustrating the rate at which alignment success decreases as sequence identity decreases. (*a*) Each point represents the average of multiple amino-acid alignments from the PREFAB version 3 set of curated alignments (Edgar 2004*b*); these reference alignments are assumed to be correct based on 3D structural super-positions of the proteins. The programs evaluated are, in approximate order of decreasing average accuracy: MAFFT v5, PSAlign, ProbCons v1, Muscle v3, T-Coffee v2 and ClustalW v1.8. The data are averaged from the experiments reported by Edgar (2004*b*), Katoh *et al.* (2005*a*) and Sze *et al.* (2006). (*b*) Each point represents the average of the simulated multiple nucleotide alignments shown in Fig. 4. The programs evaluated are, in approximate order of decreasing average accuracy: MAFFT v 5.731, ProAlign v 0.5a1, ClustalW v 1.83, T-Coffee v 3.27, Muscle v 3.52, and Clustal v V.

partly because the default gap costs in many programs have not been optimised for nucleotide sequences (as they have been for amino acid sequences), and that the Clustal, MAFFT and ProAlign programs are among the most consistent performers. As I said earlier, not much seems to have changed for multiple alignment of nucleotide sequences in the past 20 years, when most of the nucleotide-alignment procedures of Clustal and its relatives were first developed.

To further examine this idea, I have performed a comparison of several of the computer programs using the simulated nucleotide data shown in Fig. 4. I used the program settings recommended by each of the authors as being the most accurate for nucleotide sequences; and the results are summarised in Fig. 8*b*. Clearly, some things *have* changed, as ClustalW version 1.83 (from 2003) performed much better than did Clustal version V (from 1991) in the range 50–80% sequence identity. Muscle performed slightly better than ClustalW over the same range, but both it and T-Coffee performed detectably worse at <50% identity. ProAlign was very similar to ClustalW at all identities. The only consistent improvement on the ClustalW performance was by MAFFT, presumably as a result of the recent changes made to the gap costs (Katoh *et al*. 2005*b*).

There is clearly scope for further detailed evaluations of procedures for nucleotide sequence alignment, because most phylogenetic analyses involve nucleotides and therefore that is the type of alignment we need. After all, evolutionary events occur at the nucleotide level, and so this is where most of the evolutionary information exists (whereas database searching, sequence comparison and structure prediction are often better carried out at the amino acid level). Evaluations could be performed by expanding the set of structure-based RNA alignments created by Gardner *et al*. (2005), by creating some empirical nucleotide alignments of protein-coding or non-coding sequences, or by using gap-based simulation programs such as Rose (Stoye *et al*. 1998), Simulator (Fleißner 2004), DAWG (Cartwright 2005) or MySSP (Rosenberg 2005*b*). Simulated data are needed for an understanding of how variation in model parameters affects sequence-alignment procedures, but only real datasets can be used to assess which methods are best at producing the 'right answer'. What is especially needed is an evaluation that is informative with respect to the range of features that are of importance in successful homology assessment (such as BAliBASE), rather than evaluations that report the 'average' performance of the different methods across a large selection of arbitrarily chosen datasets.

Alternative strategies

The recent computerised algorithms that I have discussed here adopt one (or sometimes both) of two basic strategies to deal with potential misalignments: (i) try to avoid making the mistakes in the first place; or (ii) try to fix up the mistakes after getting an initial multiple alignment. Both types of procedure add steps to the progressive alignment algorithm, as described in the previous section, rather than replacing any of them. The program comparisons discussed above indicate quite clearly that these extra steps can result in markedly improved alignments when the patterns of sequence divergence are more complex.

Table 4. Comparison of percentage accuracy of various alignment programs for BAliBASE reference sets 1–5 and IRMBase reference sets 1–3

The percentages for the comparisons are averaged from the experiments reported for BAliBASE versions 1 and 2 and IRMBase version 1 by Thompson *et al.* (1999*b*), Katoh *et al.* (2002), Edgar (2004*b*), Riaz *et al.* (2004), Do *et al.* (2005), Subramanian *et al.* (2005), Simossis *et al.* (2005), Zhou and Zhou (2005) and Sze *et al.* (2006). For BAliBASE only the sum-of-pairs (SP) scores have been used

Program	BAliBASE					IRMBase		
	Ref1	Ref2	Ref3	Ref4	Ref5	Ref1	Ref2	Ref3
ClustalW v1	86.0	92.2	74.4	81.1	85.8	8.0	12.7	20.2
DiAlign v2	80.8	88.0	68.6	90.7	94.0	92.3	92.7	91.9
MAFFT v3	86.6	92.4	78.8	91.0	96.1	–	–	–
Muscle v3	88.6	93.7	81.8	88.3	97.5	36.2	37.8	52.3
POA v2	74.7	88.3	63.1	82.6	76.7	76.8	43.6	36.9
Praline PSI	90.4	94.0	76.4	79.9	81.8	–	–	–
ProbCons v1	90.4	94.4	83.7	91.0	97.9	66.7	68.3	77.9
PRRP	87.1	92.7	82.3	77.2	88.5	–	–	–
PSAlign	90.1	94.0	80.9	90.1	98.0	–	–	–
SPEM	90.8	93.4	81.4	97.4	97.4	–	–	–
T-Coffee v1	86.6	91.9	78.3	88.3	95.6	91.2	85.6	87.8

The programs that adopt the strategy of trying to avoid mistakes (rather than later correcting them) do so by increasing the context within which each alignment decision is made (see Fig. 9). The main limitation of progressive alignment is that it performs each alignment step in isolation, so that it never gets a global view of what the final multiple alignment will look like—it can then carry out operations that look optimal in isolation but which will not be so when viewed in a larger context. Increasing the context within which each operation is performed can then be a strategy for avoiding suboptimal decisions. In an ideal strategy, all of the pairwise alignments would contribute information to every alignment decision (i.e. the information reflects the global context), so that the multiple alignment is simultaneously optimal over all possible alignments; but this is impractical. Progressive alignment goes to the other extreme and each pairwise alignment decision is made in isolation from all other pairs, so that information is taken only from the pair of sequences directly concerned (i.e. all information reflects the local context). A compromise strategy would use information from all of the pairwise alignments that involve either of the two sequences involved in the pairwise comparison (Fig. 9).

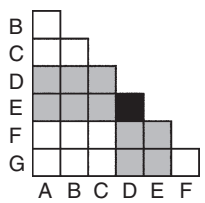

Fig. 9. Lower triangle of a pairwise comparison matrix, showing the possible comparisons used in different multiple alignment strategies for sequences A–G. For the alignment of sequences D and E, an optimal strategy would use the information from all of the comparisons simultaneously, while a standard progressive strategy would use the information only from the comparison marked in back. A compromise strategy might use the information from the boxes marked in grey as well as black.

The most successful of the programs that adopt this compromise strategy is T-Coffee (Notredame *et al*. 2000). In fact, all of the comparisons show that this program has, since its release, set the benchmark standard to which other alignment programs need to be compared. The program tries to increase the context within which the multiple alignment procedure is performed by replacing the substitution matrix with an extended library of pairwise alignments. This provides a weighting scheme based on combining different sources of information, which can then act as a position-specific substitution matrix. However, instead of using sum-of-pairs the program uses as its optimality criterion consistency among all of the pairwise alignments (Notredame *et al*. 1998). That is, the multiple alignment will be the one that is consistent with as many of the pairwise alignments as possible (Gotoh 1990). The use of the library means that the pairwise alignments can be based on information from a larger set of the sequences, rather than just the pair directly concerned.

The main advantage of this approach is that many different sets of pairwise alignments can be combined into the library and thus contribute to the multiple alignment. For example, the default strategy is to use both local (good for finding conserved motifs and blocks) and global (good for less well conserved regions) pairwise alignment strategies. Moreover, it is possible to include pairwise alignment information from any other source whatsoever, such as that based on

3D structure information if it is available for amino acid sequences (O'Sullivan *et al*. 2004; Armougom *et al*. 2006) or RNA (Bauer *et al*. 2005*a*; Siebert and Backofen 2005), thereby increasing the biological information content of the alignment. It is even possible to combine entire multiple alignments into the library, thus allowing the program to produce a consensus alignment of the alignments from other computer programs (Wallace *et al*. 2006); a similar idea exists in the ComAlign program (Bucka-Lassen *et al*. 1999). Interestingly, the extra context has also been reported to sometimes be counter-productive (Edgar 2004*b*), a result that I have also noted in some of my own alignments. Unfortunately, the program suffers from severe memory requirements, as well as the extra time taken to gather the information into the library. This means that the program is effectively limited to ~50 sequences. The PCMA program (Pei *et al*. 2003) tries to bypass this problem by aligning more-similar sequences with the Clustal algorithm and less-similar sequences with the T-Coffee algorithm, thus achieving both high quality and speed.

An alternative tactic that tries to increase the context of the alignment is to use constraints (Myers *et al*. 1996; Parida *et al*. 1999). Such methods first find locally conserved regions in the sequences and then use these as anchor sites to create the larger alignments, thus combining local and global alignment strategies. The constraints (usually gap-free conserved regions) could be defined by the user, based on prior biological knowledge such as the location of gene boundaries or functional sites, or they could be determined automatically using motif searching or some other local alignment strategy. The alignment between a consistent subset of the anchor sites then proceeds using a normal progressive global strategy. Programs that adopt this general approach include MACAW (Schuler *et al*. 1991), the various developments of DiAlign (Morgenstern 1999; Subramanian *et al*. 2005; Morgenstern *et al*. 2006), DbClustal (Thompson *et al*. 2000), FMAlign (Chakrabarti *et al*. 2004), MuSiC (Tsai *et al*. 2004; Lu and Huang 2005), RAlign (Sammeth and Heringa 2006), and Sigma (Siddharthan 2006). Align-m (Van Walle *et al*. 2004) can also be seen as fitting into this category. One of the main advantages of this approach is that it can be used to effectively deal with repeats, inversions etc, as the boundaries of the sequence blocks can be used as the constraints that anchor the alignment (Morgenstern *et al*. 2006; Sammeth and Heringa 2006). Note, also, that not all of these programs necessarily align complete sequences, as some of them allow sequence segments to remain unaligned (thus increasing specificity at the expense of sensitivity).

The second general approach to improving the progressive alignment strategy is based on the idea of refinement. That is, an initial multiple alignment is produced in some manner and is then used as the basis for an iterative series of attempts at improvement (Barton and Sternberg 1987; Corpet 1988). This is a simplified version of the procedure originally suggested by Hogeweg and Hesper (1984), in which a new guide tree was calculated from the initial alignment, leading to a new alignment, and so on. Refinement can proceed in one of several ways. The most common approach is to repeatedly divide the taxa into two subgroups and then to re-align the subgroup profiles (i.e. profile–profile alignment), until no further improvement occurs between iterations (Hirosawa *et al*. 1995; Wallace *et al*. 2005*b*). Several approaches for choosing the subgroups have been proposed, including random choice and using the internal branches of the guide tree to define the subgroups, with tree-based randomisation often considered to be the better approach (Hirosawa *et al*. 1995). An alternative successful approach is to realign each sequence individually to the profile formed by the remaining sequences (i.e. sequence–profile alignment) (Wallace *et al*. 2005*b*).

Among the earliest, and certainly the most complex, of the programs that adopt the strategy of trying to correct mistakes made early in the alignment process is PRRN (Gotoh 1995, 1996; Yamada *et al*. 2004). The iterative refinement is tree-based, and uses dynamic programming to align the profiles. However, PRRN (which now incorporates the PRRP program as well) takes this basic form of iteration one step further by also re-estimating the tree that is used to calculate the weights that define the optimality of the alignment. This creates a doubly nested iterative refinement, where the inner iterations optimise the alignment based on the current weights and the outer iterations optimise the weights, the whole process terminating when the weights have converged. This approach is limited mainly by the extra time taken for the double set of iterations, although all refinement procedures suffer from being 'hill-climbing' strategies that can get stuck in local optima rather than finding the global optimum.

Of the currently available programs that use solely primary sequence information, ProbCons (Do *et al*. 2005) is among the most successful for amino acid sequences. It optimises a consistency-based function, thus taking advantage of the increased context, but unlike T-Coffee it uses a library of pair-hidden markov models instead, thus directly modelling sequence alignment probabilities. This use of hidden markov models replaces the classification approach adopted by the other programs with an approach based on statistical modelling (Eddy 1998). After producing an initial progressive alignment the program then uses a random bipartitioning algorithm for iterative refinement, thus trying to get the best of both worlds. ProbAlign (Roshan and Livesay 2006) adopts a similar strategy but replaces the hidden markov model probabilities with partition function probabilities, with a similar degree of alignment success. Alternatively, Mummals (Pei and Grishin 2006) uses markov models that also incorporate local structural information, without making explicit structure predictions, with considerable success. Although faster than T-Coffee,

these are relatively slow compared to some of the other programs, since they have to calculate a probability matrix for all sequence pairs, and so they are best used for smaller numbers of sequences (e.g. ~five times slower than the fastest programs but ~five times faster than T-Coffee). All three programs are currently available for the analysis of amino acid sequences only.

Another iterative program worth mentioning is Praline (Heringa 1999). Among other things, it has the ability to perform a structure-based refinement in its last set of iterations for amino acid alignments. It does this by using one of several protein structure-prediction programs to predict the 3D structure of each sequence in the multiple alignment, and then in the next iteration using a different substitution weight matrix for each position depending on whether it is predicted to be part of an α-helix, a β-strand or a coil (Simossis and Heringa 2005). A similar strategy has been proposed by Jennings *et al*. (2001), although they preferred two-state structure predictions. This general approach is probably limited by the fact that automatic structure prediction currently has an accuracy of <80%. The Praline-PSI version of the program also uses as its starting point profiles of sequences gathered from database searches (using the PSI-BLAST program), rather than the original unaligned sequences (i.e. during the progressive procedure all sequence–sequence and sequence–profile alignments are turned into profile–profile alignments). This can help considerably with the alignment of sequences with low identity (Simossis *et al*. 2005), as has been shown for pairwise alignment for database searches (Ohlson *et al*. 2004) and sequence comparison (Margulies *et al*. 2006). The SPEM program (Zhou and Zhou 2005) also offers amino-acid alignments based on pre-processed sequence profiles and secondary-structure prediction, and may give superior results to Praline as it is based on a consistency measure. The main limitation of these programs is time usage, which exceeds that of all of the other programs. Zhang and Kahveci (2006) try to bypass this problem by using sequence weights based on secondary structure in a novel algorithm, which seems to produce high quality with reasonable speed.

Alternative refinement tactics also exist, all of which seek to improve a pre-existing multiple alignment. For example, Thomsen *et al*. (2002, 2003) use the alignment as the starting point for a genetic algorithm, thus seeking to move away from the local optimum that the progressive alignment has found (as a result of using only the pairwise context), while still trying to optimise the sum-of-pairs objective function. Riaz *et al*. (2005) use a tabu search in a similar manner, but use consistency as their optimality function instead. Alternatively, Manohar and Batzoglou (2005) extend the pairwise dynamic programming algorithm to groups of three sequences, which allows the local optimality of the given multiple alignment to be increased. On a different tack, Thompson *et al*. (2003) decompose the alignment into reliable and unreliable segments, and then modify the unreliable regions (once) to maximise the sum-of-pairs objective function. This attacks the problem by modifying blocks of aligned positions rather than individual sequences. Wang and Li (2004) adopt a similar strategy but use consistency as their optimality function instead, while Chakrabarti *et al.* (2006) use known conserved structural cores (e.g. from structure databases) as anchor points of reliable alignment.

A somewhat different approach is adopted by ProAlign (Löytynoja and Milinkovitch 2003). This is somewhat similar to ProbCons in that it uses a pair-hidden markov model to produce a probabilistic alignment, where sequence sites are modelled using vectors of character probabilities (including gaps). However, it combines this with the standard progressive alignment algorithm, without refinement, and an evolutionary model describing the nucleotide or amino-acid substitution process. As the alignment procedure is probabilistic this allows sampling of alternative solutions, as well as probabilistic evaluation of alignments. Thus, the posterior probability value of an aligned position can be used to identify potentially misaligned positions. This approach has been successful for aligning nucleotide sequences, probably because it uses an evolutionary model (which most other programs do not do for nucleotides), but less so for amino acid sequences, probably because its model here is less sophisticated than that of other programs. It has recently been extended, in the program Prank (Löytynoja and Goldman 2005), to treat insertions differently from deletions, which can be done since the guide tree is rooted. Insertions are then treated as individual events, whereas deletions are treated in the usual manner. This potentially leads to more realistic alignments, although the result is very sensitive to the order of progressive alignment.

Another approach is to replace the progressive alignment strategy with an alternative. For example, the PSAlign program (Sze *et al*. 2006) uses the idea of consistency among pairwise alignments as its quality measure, as above, but instead applies this only to those pairs defined by edges on a given tree. Thus, it replaces the heuristic progressive alignment step with an exact procedure that is rather more restricted. Using a minimum spanning tree results in the shortest alignment (i.e. with the minimum number of added gaps). This seems to be quite successful, although not necessarily better than iterative refinement of a progressive alignment.

In the last 10 years there have been at least 50 different methods described for multiple sequence alignment (Wallace *et al*. 2006), and I am not going to list all of them here. What I have described is intended to be no more than a brief introduction to the more relevant or commonly used of those methods. Wallace *et al*. (2006) provide an interesting classification of some of the methods that I have discussed.

Speed and genomes

In this post-genomic era it is also important to consider options for making the alignment process faster, because the speed of the algorithms becomes important as the number of sequences and / or the length of those sequences increases. The most significant point to note is that progressive alignment programs spend most of their time doing a lot of pairwise alignments for the sole purpose of producing the guide tree (Yu and Deng 2005), because most of the multiple alignment is actually produced by sequence–profile and profile–profile alignment (rather than sequence–sequence alignment). For example, for 30 taxa there are $30 \times (30–1) / 2 = 435$ pairwise alignments that need to be calculated to get the guide tree but no more than $30 / 2 = 15$ of these will then be used in the multiple alignment; and this situation becomes combinatorially worse as the number of taxa increases. So, 'scaling up' such programs to larger datasets is not trivial.

The Clustal program offers the option to speed up its production of the guide tree by calculating the pairwise distances based on *k*-tuples, as this does not require the pairs of sequences to be aligned and thus avoids the use of dynamic programming (Blaisdell 1986). This is the sort of strategy for assessing sequence similarity that is used by database-search programs such as BLAST and FASTA, and thus it is very, very fast indeed. However, it results in a much rougher guide tree, and so alignment accuracy is being sacrificed for speed. There are now several programs available that offer this strategy for increased speed but follow it by several rounds of iterative refinement (using tree-based partitioning), which is itself a speedy process (since it involves a small number of profiles rather than a large number of individual sequences). These programs combine this strategy with other speed improvements designed to deal with increasing sequence length, such as compressed alphabets, fast fourier transformation and *k*-tuple extension (Katoh *et al*. 2002; Edgar 2004*a*). Programs such as MAFFT (Katoh *et al*. 2005*a*) and Muscle (Edgar 2004*b*, 2004*c*) therefore can be extremely fast, and yet they still produce results that are comparable to the best of the other alignment programs, although there is still a significant trade-off between speed and accuracy (Katoh *et al*. 2005*b*). These are the programs recommended for high-throughput applications and larger numbers of sequences, at least for amino acid sequences. Indeed, MAFFT is the current moving target in the alignment world, with the authors seemingly determined to incorporate every new development into their program (Katoh *et al*. 2005*b*), so that it now offers more alignment strategies than most users will know how to deal with.

You may have noted that so far in this review I have said nothing specifically about aligning whole genomes. This is not because it is an unimportant topic; indeed, multiple alignment is seen as having a central role in the post-genomic era (Lecompte *et al*. 2001; Batzoglou 2005). Rather, it is because the current approach is simply to take the philosophy of the similarity programs and to upgrade their performance to deal with more data (reviewed by Pollard *et al*. 2004; Batzoglou 2005). For example, specialist programs (reviewed by Dewey and Pachter 2006) first find locally conserved regions in the genomes and then combine these anchor sites into larger (global) alignments. This is known as chaining. As an alternative tactic, one can simply identify the different genome segments independently and then align them individually using a program such as Clustal, which seems to be how most phylogeneticists deal with mitochondrial and plastid genomes.

Development of specialised algorithms for genomic sequence alignment is nevertheless considered to be a high priority (Miller 2001; Chain *et al*. 2003). This is because they have to cope with genomic re-arrangements, as discussed earlier, which confound any alignment strategy based on individual residues as single characters; this issue cannot be ignored when dealing with whole genomes, as it often can be for individual gene sequences. They also have to deal with the fact that the time requirement of the pairwise algorithms usually depends on the square of the sequence lengths, which becomes unacceptable for genome lengths. However, the main problem from the phylogeny point of view is that different genome regions may require different alignment strategies. For example, different genes may require different strategies, and coding *v.* non-coding regions almost certainly will. In this sense, the concern reflects the current interest in partitioned tree-building analyses, where different sequence partitions have different substitution models (e.g. protein-coding regions might have a codon model, RNA-coding regions have a doublet model and non-coding regions a single-nucleotide model). Any issue considered to be a problem for secondary homology assessment may also be a problem for primary homology assessment.

Nor have I yet said anything about parallel processing. Both the progressive alignment (Li 2003; Ebedes and Datta 2004; Schmollinger *et al*. 2004; Oliver *et al*. 2005) and iterative refinement (Kleinjung *et al*. 2002) procedures are candidates for parallel computing, as also are alternative alignment strategies such as genetic algorithms (Anbarasu *et al*. 2000; Nguyen *et al*. 2002) and direct optimisation as discussed in the next section (Janies and Wheeler 2001; Parmentier *et al*. 2004). This will increase their speed even further, thus making some of the more impractical methods practical.

Alternative alignment philosophies

I have noted at length that sequence alignment is primary homology assessment while tree building is secondary homology assessment. Having said this, it might then seem odd to you that in practice we treat the two things so differently—basically, we behave as a bunch of pheneticists when doing primary homology assessment, using similarity

and / or structural information to maximise the phenetic content of the sequence alignment, and then behave as a bunch of cladists when doing secondary homology assessment, constructing an evolutionary tree based on sister-group relationships. An alternative viewpoint is to treat these two procedures as two sides of the one coin, and thus adopt the same practices for both operations.

Hennig (1966) did not provide an explicit method for phylogeny reconstruction, but in his book he makes it clear that primary and secondary homology assessment are iterative procedures of reciprocal illumination. That is, we propose a primary homology and we then test it on a tree; if there is homoplasy on the tree then we must either re-assess the hypothesised homology or modify the tree. His objective for a phylogenetic analysis was to find a set of self-consistent homology statements plus a homoplasy-free tree, so that the primary and secondary homologies are congruent. This procedure was made operational in what is now known as the parsimony method of phylogenetic analysis. However, there has, in general, been little interest in re-aligning sequences to reduce homoplasy on a phylogenetic tree, so that the iterative part of his procedure has largely been lost in practice.

This idea of reciprocal illumination was formalised for sequence alignment by Sankoff *et al*. (1973), who first pointed out that we should be optimising the alignment and the tree simultaneously, since they are inter-dependent. These workers did not solve the optimisation problem in any practical sense (Sankoff and Cedergren 1983), since this problem is mathematically max-SNP-hard (Jiang *et al*. 1994), and so this philosophy was abandoned along with the proposed practice. Instead, we have treated alignment as a separate issue, which is pretty much the way a traditional phylogeneticist would do it—deciding on the homology of morphological and anatomical characters requires quite a different set of skills and knowledge compared to those required for constructing a phylogenetic tree.

For sequence alignment, however, we have adopted a series of greedy heuristic strategies to get approximations to an optimal solution in a purely mathematical sense (as described above), since this has made the alignment problem tractable (Chan *et al*. 1992). Unfortunately, there is little biological realism in these alignment procedures (i.e. they model the sequence patterns without regard to how those patterns have been formed during evolution). On the other hand, biological realism has been a much stronger focus in tree-building procedures, where increasingly complex analyses are being devised in order to make the results more relevant biologically (Morrison 2006), thus providing explicit predictions of homology and plausible explanations of evolutionary processes in addition to accurate descriptions of sequence patterns. More to the point, the assumptions applied to alignment and tree-building procedures are often not the same, an obvious example being gap penalties, which are usually treated very differently in the two analyses (Phillips *et al*. 2000). There is thus an uncomfortable dichotomy between the practice of sequence alignment and the practice of tree building.

However, there are definitely two schools of thought that have pursued the idea that tree building and alignment go hand in hand, arguing that consistency of approach for both of the steps in homology assessment (primary and secondary) is to be desired (Phillips *et al*. 2000). Hence, instead of treating alignment and tree building as unrelated activities performed sequentially, they are treated as companion activities performed simultaneously. Not as much progress has been made as their proponents would like, in terms of making the procedures practical, and this is at least one reason why our current customs have continued as they have.

One of these two schools adopts a probabilistic (sometimes called statistical) approach to tree building and therefore adopts the same criterion for sequence alignment as well (reviewed by Lunter *et al*. 2005). Explicit models of sequence evolution are constructed, usually in a likelihood context, and some criterion is then used to optimise the parameters in relation to the model, such as maximising the likelihood or maximising the bayesian posterior probability. For tree building these models have become quite sophisticated, but the progress with models for alignment since the pioneering work of Bishop and Thompson (1986) has been slow until recently. The problem with this approach to alignment has been how to incorporate indel events into the model as explicit evolutionary events. Indels are effectively ignored in the likelihood models currently used for tree building (by treating gaps as equivalent either to missing data or an extra substitution character state), but clearly they cannot be ignored for alignment. More to the point, if a contiguous set of gaps represents a single indel then it is very problematic to incorporate them into a model, as I have emphasised several times. At the moment, phylogeneticists try to bypass this problem by coding gaps as separate presence-absence characters and then analyse them using what is effectively a Jukes-Cantor model.

Nevertheless, if indels can be incorporated into the model as individual evolutionary events, then we can go straight from the sequences to the tree without the necessity of an intermediate multiple alignment. That is, the optimal result is the combined alignment and tree that maximises the likelihood (relative likelihood for a maximum-likelihood analysis and integrated likelihood for a bayesian analysis), rather than the one that separately optimises the alignment score (e.g. sum-of-pairs, consistency or log-expectation) and then the tree score (e.g. likelihood). Quite a few people have had a go at addressing this issue using likelihood models (see Lunter *et al*. 2005 for an introduction to the literature), each of them putting another individual piece into a complex

jigsaw that Thorne *et al*. (1992) characterised as 'inching toward reality'. Unfortunately, this may be over-estimating the speed of progress, as these methods are still too limited, either in terms of the size of the dataset that can be analysed (e.g. computer memory and time required) or the simplicity of the models (e.g. all gap positions are assumed to arise independently), so that none of them is yet of practical value for a realistic phylogenetic analysis. Nevertheless, two promising versions have been implemented in the AliFritz (Fleissner *et al*. 2005) and BAli-Phy (Redelings and Suchard 2005) computer programs, along with POY (Wheeler 2006).

Note that this is an explicit attempt to put some biological insight into sequence alignment, by building mathematical models derived from the biological processes of evolution. Most of the alignment algorithms mentioned in previous sections make little or no reference to any underlying evolutionary model (Thorne and Churchill 1995). Unfortunately, even the approach discussed here still uses evolutionary models that are quite limited. For example, other mutational events that affect blocks of sequence simultaneously, such as inversions, translocations, transpositions and tandem duplications, have not yet been incorporated into the models, which will be a major challenge. Similarly, models that incorporate non-independence among characters have not yet been employed, such as codon models for protein-coding sequences (allowing gaps of three positions) and doublet models for structural RNAs (allowing for base pairs in helices). This means that different types of sequences (e.g. protein-coding, RNA-coding, non-coding) cannot yet be analysed with different models, which will probably be necessary for a comprehensive phylogenetic analysis. Furthermore, in the probabilistic context it is perhaps anomalous to be pursuing a single multiple alignment (i.e. a point estimate), as this ignores the large set of almost-equally optimal alignments, which leads to biases in the parameter estimates and tree probabilities. A use of bayesian posterior probabilities may be a more effective strategy (Allison and Wallace 1994; Zhu *et al*. 1998) as this automatically provides an assessment of reliability. Finally, it has long been recognised (Allison *et al*. 1992) that selection pressure is missing from all of our alignment and tree-building models (although see Hein and Støvlbæk 1996), which biases the estimates of the rates and types of substitutions occurring (e.g. accounting for synonymous *v.* non-synonymous substitutions).

The other school of thought adopts the parsimony principle for tree building and thus also adopts the same criterion for sequence alignment (Vingron 1999; Wheeler 2001*a*, 2001*b*, 2005). This was actually the approach adopted by Sankoff *et al*. (1973). Here, 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. This makes the overall phylogenetic approach philosophically consistent, as the parsimony criterion is used to evaluate both the (implied) alignment and the (explicit) tree. However, it also makes the optimisation problem extremely difficult, if not impossible, because the optimisation has to occur simultaneously over all possible alignments and all possible trees, and even parsimoniously reconstructing indels on a given tree is NP-hard (Chindelevitch *et al*. 2006). Thus, a heuristic approach is needed in practice, the first being provided by Hein (1990) who iterated between evaluation of the alignment and construction of a distance-based tree (trying to minimise the edge lengths of the tree). This approach has subsequently been developed by Vingron and von Haeseler (1997), Schwikowski and Vingron (1997*a*, 1997*b*, 2003), Lancia and Ravi (1999) and Trystram and Zola (2005) under the name generalised tree alignment. Alternatively, Wheeler and Gladstein (1994) provided a more direct parsimony implementation by performing an iterative heuristic search through tree space for the alignment that produces the minimum-cost tree.

However, these procedures simply use tree construction / search to produce a multiple alignment. Wheeler (1996, 1998, 2002, 2003*a*) has devised heuristic methods (under the name direct optimisation) that bypass the need to produce a separate multiple alignment at all, by directly optimising ancestral sequences while treating gaps as a fifth character state rather than as missing data, thus moving closer to the likelihood methodology described above. This approach has been extended by Wheeler (1999, 2003*c*) to treat a contiguous series of gaps as a single evolutionary event (i.e. an indel), but not yet explicitly the other sequence-block events. These methods are all implemented in the POY computer program. None of them have yet been tested thoroughly (unlike the progressive alignment strategies), although there have been investigations of the effect of varying parameters such as gap costs and fragment size (Cognato and Vogler 2001; Giribet 2001, 2002; Petersen *et al*. 2004; Aagesen *et al*. 2005; Terry and Whiting 2005). (Note that these parameters are estimated in the probabilistic models above rather than fixed, whereas here substitution and gap costs are set so as the maximise secondary homology.) Other issues that need addressing include: the fact that the parsimony score is only an approximation (Shull *et al*. 2001); potential problems with sequence blocks (Lee 2001); and quantifying uncertainty in the phylogenetic tree, since methods such as bootstrapping are inappropriate due to that fact that the aligned characters are no longer independent (Redelings and Suchard 2005).

Note that the argument for an alignment-free tree-building method, by either the statistical or parsimony approach, is at the heart of the philosophical difference from current practice. Both the probabilistic and parsimony methods

are applied to directly derive a phylogenetic tree from the individual sequences, thus effectively bypassing the need to produce a separate multiple alignment at all. 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. If a multiple alignment is explicitly needed then it can be derived by aligning the sequences onto the tree (i.e. by inferring the sequences of the ancestors), a procedure known as implied alignment (Wheeler 2003*b*; Giribet 2005) or posterior-decoded or most-likely alignment (for the probabilistic methods). Note that this alignment will depend on the tree (i.e. a different tree will imply a different alignment), which in turn will depend on the specific sequences and regions included in the dataset (see the example in Shull *et al*. 2001).

From this perspective, current practice is seen to add an unnecessary step to the process of constructing a phylogenetic tree (i.e. a separate multiple sequence alignment). Furthermore, this extra step is performed by a type of analysis that has been co-opted from the sequence-comparison literature rather than being specially developed for phylogenetic analysis. (Note, the sequence-comparison people think that ideally the phylogenetic tree should come first, so that they can derive an evolutionary-based multiple alignment from it, as this is a useful thing for database searching and structure prediction.) Therefore, not only are we using an inappropriate methodology for sequence alignment, but we don't actually need to be using it at all (De Laet 2005). Moreover, these procedures are designed to deal explicitly with the fact that similarity = homology + analogy, by integrating primary and secondary homology assessment. None of the traditional alignment procedures do this, since they rely entirely on pattern matching.

The alternative viewpoint is that historically homology assessment and tree construction have been treated as separate issues, with good reason, and so it doesn't seem strange to treat sequence data in the same way today. That is, we have always used comparative biology to produce our hypotheses of homology and then tested these hypotheses on a tree. Thus, we have kept hypothesis generation separate from hypothesis testing, which matches our historical separation of primary and secondary homology assessments (Simmons and Ochoterena 2000). However, because we are using the same data to generate the hypotheses and to test them, this cannot be seen as a valid test in either a philosophical or a statistical sense. For example, all alignment methods shuffle character states among characters as they proceed (as noted earlier) but they do so with the implicit objective of defining the characters (the columns in the alignment). Shuffling the character states while building the phylogenetic tree means that congruence among the characters becomes part of the definition of the characters rather than a test of them. If nothing else, this can create artefacts as a result of the inter-play of alignment and tree building.

Perhaps one of the most obvious objection to combining primary and secondary homology assessment is that it weakens the independence of different datasets when they are combined, because the tree topology supported by one dataset (e.g. the gene tree for a locus) can influence the alignment of another dataset (Simmons 2004). This second dataset is then not being used to independently test the tree supported by the first dataset but is instead merely being assessed for its degree of congruence with that tree.

However, the biggest conceptual issue from the perspective that I have been pursuing in this paper is that of homology. We can combine alignment and tree building provided they both recognise that character homology is at the heart of phylogenetic analysis. The parsimony approach is usually considered to do this (Haszprunar 1998) but it is not clear that the likelihood models do so. Indeed, both methods compound the confusion between descriptive and ontological parsimony alluded to earlier, because this methodological criterion is now applied to both character definition and character testing. This leads to the situation where 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). This is a form of indirect character and character-state weighting.

From the practical viewpoint, it is unlikely that either the parsimony or the model-based alignments will be what we actually want. That is, the maximum-parsimony alignment may be philosophically justifiable and the maximum-likelihood alignment may be statistically justifiable, but it is improbable that either will represent the true evolutionary alignment. When a comparison is made to structure-based alignments, neither the parsimony nor the model-based alignments are closer than are the common heuristic alignment algorithms (Morrison and Ellis 1997; Miklós *et al*. 2004; Gillespie *et al*. 2005*a*), although the parsimony and model-based approaches certainly add more gaps to the multiple alignment than do the progressive-alignment programs (Morrison and Ellis 1997; Whiting *et al*. 2006). Thus, combining two mathematical optimality problems into one does not necessarily get you any closer to biological reality. There have also been several empirical comparisons of alignments from the POY and Clustal programs. Unsurprisingly, those people who used the parsimony score as the assessment criterion found that the parsimony alignment was best (Giribet *et al*. 2002; Wheeler 2003*b*; Terry and Whiting 2005; Whiting *et al*. 2006), and those who used the likelihood criterion found that the model-based alignment was best (Whiting *et al*. 2006). This is uninformative because Clustal does not try to optimise either score. Clearly, what is needed is an

independent assessment criterion, such as congruence with taxonomy (e.g. Shull *et al*. 2001; Belshaw and Quicke 2002; Laurenne *et al*. 2006).

Biological sequence alignment

A result obtained from any computational method is unlikely to be the ultimate answer to a particular biological inference—the method is unlikely to give the mathematically highest score to the biologically correct result under all circumstances. In the previous sections I have repeatedly emphasised this point with regard to multiple sequence alignment for phylogenetic purposes. Without knowledge about the processes that generated the patterns of residue variation, alignment cannot be accurate. Indeed, given the fact that phylogeny deals with historically unique events, maybe it is illogical even to think that there could be a general method for phylogenetic analysis. Each dataset may need to be taken on its own merits, with a unique approach adopted depending on the specific interaction that has occurred between data and history.

Even if we do adopt a mathematical approach, most of the commonly used multiple-alignment programs are not based on any evolutionary considerations at all, but simply use some mathematically tractable algorithm for maximising sequence similarity. Their success for our purposes is thus predicated solely on the extent to which similarity = homology. That they 'work' at all probably tells us quite a lot about the role of analogy in evolutionary history, or at least our perception of it.

Moreover, most of the recent attempts to improve these programs apply only to amino acid sequences. Amino acids come in a wide range of flavours, with distinct patterns of physico-chemical properties, which means that there is more information to extract and use in developing alignment strategies. This is not so for DNA sequences, and so the scope is more limited. Even worse, gaps in RNA-coding and non-coding (e.g. with regulatory motifs) sequences often appear to be more haphazard than they are for protein-coding sequences, with the nucleotides not lined up in neat columns the way amino acids often are. It is therefore unsurprising that the developers of alignment programs have found DNA alignment to be a nuisance (Higgins *et al*. 2005).

The development of statistical alignment and direct optimisation techniques is one possible response to this general situation, where the evolutionary and philosophical models used for constructing phylogenetic trees are extended to include sequence alignment as well. It is not yet clear how productive these attempts will be, but obviously they are unlikely to succeed to any greater extent than they do for tree building alone. We certainly have a long way to go if we want any of the models to be biologically realistic.

So, quality control in multiple sequence alignment for phylogenetic purposes relies entirely on the biological insight of the scientist. The biologist needs to control the alignment process at all stages to make sure that the final alignment represents a series of plausible hypotheses of homology. This does not require an entirely manual alignment strategy, but it does seem to imply manual quality control at least. In this sense, it has been suggested that we should only use the mathematical algorithms as heuristic procedures to produce a first approximation for the final multiple alignment (Higgins *et al*. 1996; Poch and Delarue 1996). This section of the review discusses how we might go about doing this.

Objectivity and reproducibility

It is a basic tenet of science that the component activities should be objective and reproducible. So, if phylogenetic analysis is to be a part of science then the process by which we obtain a multiple sequence alignment must, itself, be objective and reproducible. One of the most common arguments against manual intervention in sequence alignment has been that it is not objective, in the sense that there is no explicit protocol to describe how it is done, and it is not reproducible, in the sense that different researchers are likely to get different results when aligning the same set of sequences. Clearly, I am not going to suggest alignment strategies that can be subject to either of these objections. I therefore need to be explicit about what objectivity and reproducibility mean in the context of a multiple sequence alignment.

Here, objectivity simply refers to the scientist's ability to adequately describe and justify the criteria being used to make decisions. An automated procedure is usually considered to be objective rather than subjective because the decision-making activities have been incorporated into an algorithm, and the scientist has no direct say in the subsequent decisions. Unfortunately, this attitude begs the question, because all that has happened is that the subjective decision has been moved back one step, so that it exists in the choice of an algorithm in the first place rather than in the choice of which residues to align.

That is, there is no generally accepted protocol for sequence alignment, and no apparent objective means of choosing among the available protocols. The decision as to which computer program (i.e. algorithm) to use is thus a subjective one. More to the point, few people seem either to describe how the choice was made or to justify the criteria used for that choice. Given that it is well known that the choice of algorithm can have a major effect on the result (and this effect increases as sequence identity decreases), computerised sequence alignment is currently, in practice, not particularly objective. Potential subjectivity in computerised alignment is dealt with by convention rather than by objectivity.

Reproducibility has two different components, which should not be confounded: (i) imitation, which is the ability to ape a specified series of steps to arrive at an identical

conclusion; and (ii) repeatability, which is the ability to independently arrive at the same conclusion irrespective of the precise series of steps. Scientific evidence is predicated upon (ii), since independent replication is seen as the strongest form of evidence, while computers are good at (i) only, without the intervention of the user.

The key to (i) is a precisely detailed set of instructions, such as a laboratory protocol or a computer program. However, whereas a laboratory protocol usually leaves considerable leeway for adjustment by not being too specific about many of the details, a computer program can only imitate itself time after time. The key to (ii) is an explanation of the objective and how it has been implemented, so that we can think about it and independently execute it. However, a computer program can easily be subject to unthinking use, literally at the press of a button. Imitation is often held up as a positive feature of computing, which it is, but the problem is that it is not a substitute for repeatability, which is what a scientist needs.

The important distinction, then, is not between 'manual' alignments and 'automated' alignments, as it is often presented in the literature. Both types of alignment can fail the criteria of objectivity and repeatability and both can meet them. A good recent example of a carefully reasoned manual assessment of a multiple alignment is presented by Lebrun *et al*. (2006), in which the authors provide a meticulous structural and functional analysis of a difficult alignment problem before phylogenetic analysis.

As far as computerised alignments are concerned, they add little to science if imitation is their only utility. For example, in this age of computerised databases there is little practical purpose to being able to reproduce someone else's alignment by re-running the same program, since it could be stored in a database. More to the point, thousands of people could use the same program and yet get different results, because (almost all of) these programs have parameters that can be adjusted. From this perspective, it is not the use of the computer program itself that makes the result reproducible but the use of the same parameter values (cf. Kjer *et al*. 2006). The main difference from manual alignment is simply the ease with which the alignment protocol can be described (and also carried out). Unfortunately, the computerised protocols are rarely specified in enough detail for repeatability, because we are often not told which version of the program was used nor what were the parameter settings, and sometimes not even what program was used.

A simple example will suffice to emphasise the need to specify all three details (program, version, parameters) in order to be able to reproduce a computerised analysis. Figures 3 and 4 of Morrison and Ellis (1997) show the effect on a particular alignment of varying the gap-opening (GOP) and gap-extension (GEP) parameter values when using ClustalW version 1.5. Included in the display are the results from the supposed default parameter values (GOP = 10, GEP = 0.5). However, the results shown are not those derived from simply accepting the default values (i.e. by changing nothing before running the program), but are instead those derived by manually changing the parameter values (i.e. manually changing the parameter settings to the default values). This is because the two alternative procedures did not produce the same alignment. This 'feature' is, fortunately, absent from the current release of ClustalW, where automatically and manually choosing the default values *do* produce the same result.

It is therefore worth highlighting that perhaps the most under-appreciated aspect of different version numbers of alignment programs is possible changes to the default GOP and GEP, which are often changed substantially between versions (e.g. Katoh *et al*. 2005*b*). This can have a quite dramatic effect on the resulting alignment (indeed, that is often the main reason for making the change), so that alignments are only reproducible if the exact version number of the program is specified.

For manual intervention in alignments, we need to have explicit criteria for the procedures being used, so that we can clearly describe those procedures along with some quantitative measure that tells us whether our alignment is 'improved' by the intervention or not. A minimum standard in science is that sufficient information should be provided for an independent experimenter to be able to repeat the work, and this applies to all forms of sequence alignment as well (Henikoff 1991). We do not necessarily need the minute details of each individual adjustment made, but we do need a description of why and how the adjustments were carried out. We also need a copy of the final alignment so that anyone who attempts to repeat the procedure can compare the previous results to their new ones. To this end, all phylogenetic multiple alignments should be publicly available; for example, the alignment can be deposited in one of the available database repositories such as EMBL-Align (Lombard *et al*. 2002), PopSet (Brawley 1999) or TreeBASE (Morell 1996).

Unfortunately, there are several reasons why sequences cannot be accessed freely in databases, even if they have been submitted. For example, the authors (= owners) can usually specify release dates, and can keep modifying those dates. Also, they often have the ability to delete information after it has been released. They may also not provide the correct reference information, or that information may subsequently be changed. For example, Lawrence *et al*. (2002) refer to three databased amino-acid sequence alignments (EMBL DS43278, DS43279, DS43280) for their 137 kinesin sequences. However, none of these alignment numbers can currently be accessed in the EMBL database. Instead, the alignments can be accessed in that database under the numbers ALIGN_000356, ALIGN_000357 and ALIGN_000358, as indicated by Lawrence *et al*. (2004). Unfortunately, the latter two alignments are identical, while

the first alignment has slightly different taxon names and almost always shorter sequence lengths than the other two alignments (except for CelU61947, which has a completely different sequence), none of which matches the published descriptions. Alternatively, Kroken and Taylor (2001) quote a TreeBASE number (SN376-1131) for their sequence alignment. However, 'SN' numbers are temporary submission numbers, and cannot be used by other people to access the data. Failure to get a permanent 'S' study accession number means that the data remain unreleased and therefore inaccessible to others. Laurenne *et al*. (2006) quote the TreeBASE Referee PIN instead of the study accession number, thus making the data unavailable in a normal search.

Having the alignment, exactly as analysed by the authors, available as supplementary material on the journal's web page should therefore be considered essential. This allows the reader to check the alignment and, if necessary, re-interpret the authors' conclusions in the light of any apparent discrepancies or inconsistencies. One example is the alignment of the 23S (large subunit) rDNA sequences associated with the paper by Badger *et al*. (2005) about possible horizontal gene transfer in α-proteobacteria. In the journal's online data file there are 2133 aligned positions, but gapped positions have been removed and so the full alignment cannot be reconstructed. Nevertheless, there are clear misalignments in several places, usually where a long motif in one of the sequences is not aligned against the identical motif in the other sequences. These require a single-nucleotide gap to be placed at either end of the motif in order to effect each re-alignment. Unfortunately, these re-alignments affect 10 of the 18 species in the dataset; and they have consequences for the tree building needed for the authors' arguments.

Alignment within context

It may seem a bit trivial to say it, but it is important to point out that every phylogenetic multiple alignment exists only within its own immediate context. That is, if either the character sampling or the taxon sampling is changed then the alignment is likely to change. This is because we are trying to reconstruct historically unique events, and our ability to do so depends entirely on the information contained in the particular dataset at hand. A straightforward way to add biological insight to a sequence alignment is thus to carefully plan the framework within which the alignment will take place.

A simple example to illustrate this point is shown in Fig. 10. For these 10 intron sequences there is a complex gap structure near position 20, consisting of two separate gapped blocks. This alignment structure is dominated by the presence of *Spartina bakeri*, which is the only sequence that spans both blocks. If this species is removed then the two gaps are concatenated by the alignment algorithm. Thus, two different sets of hypotheses of homology emerge, depending on whether *S. bakeri* has been sampled or not. This occurs because the optimisation procedure used by the computer program only produces a result that is optimal for the local context—there is no such thing as a universal optimal alignment, only an alignment that is optimal for the particular dataset being analysed.

The effect of taxon sampling on sequence alignment cannot be over-stressed (Simmons and Freudenstein 2003; Simmons *et al*. 2004). It is therefore always recommended that taxon sampling be increased to whatever extent is practicable, as this reduces the chance that some of the sequences will have no near relatives and thus be hard to align (i.e. it has the same effect as breaking up long branches in a tree-building analysis). Sampling only selected exemplars is often seen as an acceptable strategy in phylogenetic analysis, but this approach should not be taken lightly for sequence alignment.

To this end, a strategy commonly used in the sequence-comparison and database-search literature is to construct sequence profiles, thus employing profile–profile alignment rather than sequence–sequence alignment. The increased context of the profiles can allow sequences with very low similarity to be aligned, because the other sequences in the profiles form a series of linking intermediates in terms of similarity. This approach is built into the MAFFT and Praline computer programs, as discussed above. The sequences for the profiles are obtained by database searching, and are chosen solely for their similarity to the sequences

```
                          1       10        20        30        40
Spartina alterniflora     GTGAGCCTATTTCCGTCCG---TTGCCGTATTGGGAGGGTT
Spartina maritima         GTGAGTCTATTTCCGTCCG---TTGCCGTATTGGGTGGGTT
Spartina foliosa          GTGAGCCTATTTCCGTCCG---TTGCCGTATTGGGAGGGTT
Spartina argentinensis    GTGAGTCTGTTTTCGTCCG---TTGCCGTGTTGGG-GGGTT
Spartina densiflora       GTGAGTCTATTTTC----G---TTGCCGTATTGGG--GGTT
Spartina cynosuroides     GTGAGAGCCTGCAT----GAGCTTGTCGT-TTGGG--GTTT
Spartina arundinacea      GTGAGAGCCTGCAT----GAGCTTGTTGT-TTGGG--GTTT
Spartina patens           GTGAGAGCCTGCAT----GAGCTTGTCGT-TTGGG--GTTT
Spartina bakeri           GTGAGAGCCTGCATGCATGAGCTTGTTGT-TTGGG--GTTT
Spartina pectinata        GTGAGAGCCTGCAT----GAGCTTGTCGT-TTGGG--GTTT
```

Fig. 10. Beginning of the ClustalW (v1.83 with default settings) alignment of intron 8 of the granule-bound starch synthase (*Waxy*) gene of ten species of *Spartina* (Poaceae). The *S. bakeri* sequence provides the context for the alignment of the remaining sequences. The nucleotide data are from Baumel *et al.* (2002).

being aligned (i.e. their taxonomic relationship to the study sequences is not taken into consideration). These database sequences only play a part in the sequence alignment, and are discarded again after the alignment has been constructed. This strategy has apparently not been adopted in phylogenetic analyses, but it may pay dividends for sparsely sampled taxonomic groups.

Constrain the alignment

One of the simplest ways to add biological insight to a progressive sequence alignment is to constrain various aspects of the procedure so that they conform to pre-existing biological knowledge. This prevents artefacts of character and taxon sampling from dominating the final alignment, and allows the user's personal knowledge of the dataset to over-ride the generic 'conventional wisdom' embodied in computer programs (Myers *et al*. 1996). The two most obvious features to constrain are the order in which the sequences are aligned and the presence of conserved sequence blocks (e.g. motifs). Both of these constraint types can be implemented in a semi-automatic manner.

Aligning sequences in the order specified by a pre-existing taxonomic scheme is a long-standing suggestion (Mindell 1991). For the procedure to be objective and repeatable all that is required is a description of the sequence groups that are to form the initial profiles and the source of the taxonomy / phylogeny. For example, Page (2000) used ClustalX to align 225 domain III 12S rRNA sequences of insects, but instead of using an automatically generated guide tree he constrained the progressive procedure to match a putative phylogeny. That is, taxonomic groups of sequences (e.g. termites) were aligned separately, and then these small alignments were combined using profile–profile alignment in the order specified by the phylogenetic tree. Similarly, Pettersson *et al*. (2005) used ClustalW to align 179 glutathione *S*-transferase amino-acid sequences, but constrained the order of alignment to match the known structural classes of the enzyme. That is, sequences were first aligned within each of the 12 classes, and then these small alignments were combined using profile–profile alignment in the order specified by previous studies of the phylogenetic history. The newly acquired sequences were then added at the end, using sequence–profile alignment.

Some of the progressive-alignment computer programs will allow the user to control the order of alignment, notably the Clustal programs. Unfortunately, many others offer no such facility. It is also important to note that genetic similarity may be a better criterion than sister-group relationships for defining constraints, as the latter may result in quite dissimilar sequences being aligned if the relationship involves long branches on the phylogenetic tree (Edgar 2004*c*).

The concept of constraining sequence blocks was discussed in an earlier section. In this strategy, sequence blocks that are conserved across most (or all) of the sequences are used as anchor points between which automatic alignment can occur. If the constraints are defined by the user based on prior biological knowledge, then the rationale for identifying such sites needs to be made explicit (e.g. the location of gene boundaries, functional sites or structural features), if the procedure is to be objective and repeatable. Alternatively, if the constraints are determined automatically, such as by using computerised motif searching or some other local alignment strategy, then the biological foundation needs to be carefully thought out, if there is to be any increase in biological insight.

An example is provided by the work of Titus and Frost (1996). These authors used the MALIGN computer program (Wheeler and Gladstein 1994) to align the mitochondrial 12S rRNA and valine tRNA sequences from their group of 10 lizard species. However, they constrained the alignments so that the stem regions of the sequences (i.e. the double-stranded parts of the helices) were aligned to one another, while allowing sequence similarity to determine the optimal placement of gaps in the other regions of the sequences. The stem regions were identified using published secondary-structure models, based on positional similarity, base-pair complementarity and compensatory substitutions. Of the 1129 positions in the alignment, 457 were constrained, based on 35 helical regions. Their argument for proceeding in this manner was that there is no extrinsic model for postulating positional homology in non-stem regions, and so mathematical optimisation is appropriate for these parts of the sequences. Similar examples are provided by Shull *et al*. (2001) and Giribet (2002), although Shull *et al*. (2001) used sequence motifs as well as secondary structure to define the constraints.

A range of computer programs are available that directly implement constraint alignment, as listed in an earlier section, although not all of these allow user-specified constraints. Also, some alignment workbenches allow sequence blocks to be manually anchored and thus not subject to automatic re-alignment, notably BioEdit, ClustalX, PAT and MACAW.

Staggered alignment

Multiple sequence alignment for phylogenetic purposes is all about producing hypotheses of potential homology, where each aligned position represents a set of hypothesised evolutionary events. Thus, 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. Unfortunately, most computerised alignment procedures do not do this, because the parsimony principle leads them to compress gaps as much as possible, leading to over-lapping non-homologous sequences (sometimes called over-alignment). Therefore, perhaps the simplest way to increase the biological insight of a multiple sequence alignment is to apply this clear-cut dictum manually.

This issue is related to the distinction between sensitivity (or power) and selectivity (or confidence) in a data analysis. Here, sensitivity refers to the ability to detect all of the residues that should be aligned (i.e. no false negatives) and selectivity refers to the ability to align only those residues that should be aligned (i.e. no false positives). Assessment of the accuracy of sequence alignment procedures has concentrated on sensitivity rather than selectivity (Lambert *et al*. 2003), given that there is a trade-off between the two (increasing one decreases the other if the sample size remains constant). This means that falsely aligning non-homologues has not been seen as a problem that needs addressing, and hence users must deal with this for themselves.

The basic dilemma appears to be that keeping non-homologous residues unaligned creates alignments that do not match most biologists' idea of what an alignment should look like. In fact, this form of alignment has been referred to as a staggered alignment (Barta 1997), because the residues within gapped regions no longer form neat columns but are offset with respect to each other. A similar issue has been identified for alignments of protein structures (Marsden and Abagyan 2004), where some methods 'over-align' compared to others.

An illustrative example is shown in Fig. 11, where two versions of the same HIV-1 sequence region are juxtaposed. The gapped part of the alignment, presumably

Original alignment

```
       1       10        20        30        40        50        60        70
 C.1   TCAGAACAGAAC------------------------------AGAGCCAACAGCCCCACCAGAAGAGAG
 C.2   CCAGAACAGATCAGAGCCAGCAGCCCCAAC-------------AGTACCAACAGCCCCACCAGCAGAGAG
 C.3   TCAGAGCAGACCAGAGCCAACAGCCCCACCAGAGAGTCTCAGACCAGAGCCAACAGCCCCACCACCAGAGAG
 C.4   CCAGAGTAGACC------------------------------AGAGCCAACAGCTCCACCAGCAGAGAG
06.1   TCAGAACAGGCC------------------------------AGAGCCAACAGCCCCACCCATAGAGAG
06.2   TCAGAACAGGCC------------------------------AGAGCCAACAGCCCCACCCGCGGAGAG
06.3   TCAGAACAGGCC------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
06.4   CCAGAACAGGCCAGAACAGAACAGGCC---------------AGAACCCTCAGCCCCACCTGCAGAGAG
12.1   TCAGAACAGGCC------------------------------AGAACCAACAGCCCCACCAGCAGAGAG
12.2   TCAGAACAGGCC------------------------------AGAGCCAACAGCCCCGCCAGCAGAGAG
12.3   TCAGAACAGACC------------------------------AGAGCCAACAGCCCCGCCAGCAGAGAG
12.4   TCAGAACAGGCC------------------------------AGAGCCAACAGCCCCGCCAGCAGAGAG
13.1   TCAGAGCAGACCAGGACCAACAGCCCCACCAGAGAGCAGACC---AGAGCCAACAGCCCCACCAGCAGAGAG
13.2   TCAGAGCAAACCAGGGCCAACAGCCCCACCAGAGAGCAGACC---AGAGCCAACAGCCCCACCAGCAGAGAG
14.1   CCAGAACAGGCC------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
14.2   CCAGAACAGGCC------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
14.3   CCAGAACAGGCC------------------------------AAAGCCAACAGCCCCACCCGCAGAGAG
14.4   CCAGAACAGGCT------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
 O.1   GCAGAGACCAGC------------------------------ACACCCATCAGCCCCACCGATGGAGGA
 O.2   GCAGAAACAAGT------------------------------GTCCCCATCAGCCCCACCAATGGAGGA
 O.3   GCAGAGACAAGT------------------------------GTCCCCATCAGCCCCACCGATGACGGA
 O.4   ACAGAGACAAGT------------------------------GTCCCCATCAGCCCCACCGATGACGGA
 N.1   CCAGACAACAACAAGGAA------------------------AGAGCCCACAGCCCCGCCACTAGAGAG
 N.2   CCAGACAACAACAAGGAG------------------------AGAGCCCACAGCCCCGCCACTAGAGAG
```

Homology alignment

```
       1       10        20        30        40        50        60        70        80        90
 C.1   TCAGAACAGAAC------------------------------------------------------AGAGCCAACAGCCCCACCAGAAGAGAG
 C.2   CCAGAACAGATCAGAGCCAGCAGCCCCAAC-------------------------------------AGTACCAACAGCCCCACCAGCAGAGAG
 C.3   TCAGAGCAGACCAGAGCCAACAGCCCCACCAGAGAGTCTCAGACC---------------------AGAGCCAACAGCCCCACCACCAGAGAG
 C.4   CCAGAGTAGACC------------------------------------------------------AGAGCCAACAGCTCCACCAGCAGAGAG
06.1   TCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCACCCATAGAGAG
06.2   TCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCACCCGCGGAGAG
06.3   TCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
06.4   CCAGAACAGGCC------------------------------AGAACAGAACAGGCC------AGAACCCTCAGCCCCACCTGCAGAGAG
12.1   TCAGAACAGGCC------------------------------------------------------AGAACCAACAGCCCCACCAGCAGAGAG
12.2   TCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCGCCAGCAGAGAG
12.3   TCAGAACAGACC------------------------------------------------------AGAGCCAACAGCCCCGCCAGCAGAGAG
12.4   TCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCGCCAGCAGAGAG
13.1   TCAGAGCAGACCAGGACCAACAGCCCCACCAGAGAG---CAGACC---------------------AGAGCCAACAGCCCCACCAGCAGAGAG
13.2   TCAGAGCAAACCAGGGCCAACAGCCCCACCAGAGAG---CAGACC---------------------AGAGCCAACAGCCCCACCAGCAGAGAG
14.1   CCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
14.2   CCAGAACAGGCC------------------------------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
14.3   CCAGAACAGGCC------------------------------------------------------AAAGCCAACAGCCCCACCCGCAGAGAG
14.4   CCAGAACAGGCT------------------------------------------------------AGAGCCAACAGCCCCACCCGCAGAGAG
 O.1   GCAGAGACCAGC------------------------------------------------------ACACCCATCAGCCCCACCGATGGAGGA
 O.2   GCAGAAACAAGT------------------------------------------------------GTCCCCATCAGCCCCACCAATGGAGGA
 O.3   GCAGAGACAAGT------------------------------------------------------GTCCCCATCAGCCCCACCGATGACGGA
 O.4   ACAGAGACAAGT------------------------------------------------------GTCCCCATCAGCCCCACCGATGACGGA
 N.1   CCAGACAACAAC------------------------------------------------AAGGAAAGAGCCCACAGCCCCGCCACTAGAGAG
 N.2   CCAGACAACAAC------------------------------------------------AAGGAGAGAGCCCACAGCCCCGCCACTAGAGAG
```

Fig. 11. Two alternative alignments of part of the *gag* gene from seven HIV-1 subtypes, with 2–4 sequences per subtype. The second alignment aligns only homologous positions, thus illustrating the concept of a staggered alignment. The nucleotide data are from the 2001 HIV-1 Subtype Reference Alignment of the HIV Sequence Database (http://hiv-web.lanl.gov/content/index; accessed 19 November 2006).

representing insertions in some of the sequences, seems to be phylogenetically informative, at least for some of the HIV-1 subtype groups. The alignment of the insertions is straightforward within each subtype, but is not always so between subtypes. For example, the alignment of subtype 13 against subtype C is unproblematic (and seems to indicate an historical association), but the alignment between these subtypes and either subtypes 06 or N is somewhat arbitrary.

Thus, a more plausible alignment is presented in the second part of the figure, where only putatively homologous nucleotides are aligned. In this case, three independent indel events are postulated for the different subtypes (one in subtype N, one in 06, and one in C + 13), followed by two independent indels within subtype C, to create the complete gapped region. This alignment also maintains the reading frame of the amino acids, as all of the gaps are in multiples of three nucleotides. (Note that there is little evidence here for the monophyly of subtype C.)

The matter of the non-neat alignments that result from not aligning indel regions is discussed at more length by Higgins *et al*. (2005), who suggest that a visual paradigm shift is needed. Our current perception of alignments started when automated procedures were developed in the late 1980s; previous manual alignments were frequently staggered (e.g. the well studied alignment of Kreitman 1983), but they have rarely been employed since then (e.g. Hancock and Vogler 2000; Kelchner 2000; Sanchis *et al*. 2001). The only computer program currently available that treats insertions as unique events that should not be aligned against other sequence blocks is Prank, as discussed in a previous section.

Thus, in practice staggered alignments may need to be produced by manual re-alignment of a pre-existing alignment. If so, then some criterion needs to be specified for how non-homology is to be recognised. Perhaps the simplest criterion is to identify indel regions where there is no clear similarity between the sequences, as I did in the example above—unaligning such sequences is then a conservative procedure (i.e. homology is only postulated when the primary homology assessment is clear, so that 'I don't know' is treated as equal to 'not homologous'). The objective is that our primary homology assessment should result in an alignment that explicitly indicates both potential homology (aligned residues) and non-homology (unaligned residues). Our hypotheses of homology can then be assessed on a phylogenetic tree, for confirmation or rebuttal.

Criteria for manual realignment

Several literature surveys (e.g. Whiting *et al*. 2006; Kjer *et al*. 2006) have suggested that the most commonly reported procedure for multiple sequence alignment is to use a computer program and then to follow this with manual readjustment (i.e. 'by eye'). As I have argued, this can be an effective way of introducing some biological (as opposed to bioinformatic) insight into the resulting alignment. Moreover, formal comparisons with the results of other alignment strategies have indicated that this approach is not necessarily any less accurate than fully computerised procedures (e.g. Sanchis *et al*. 2001; Giribet 2002; Whiting *et al*. 2006; Kjer *et al*. 2006). However, this can only be considered to be a scientific procedure if it is objective and repeatable. This involves having (a) a description of the manual process that is detailed enough for someone else to be able to repeat it, and (b) a quantifiable criterion for determining whether the alignment quality has been improved by the procedure. Here, I discuss various ways in which both of these conditions might be met.

The informal objective of manually scanning an alignment is to look for 'problems' that need fixing. Our minds have a simultaneous overview of the alignment that is not available to the computer programs, especially those based on a progressive strategy, and thus we expect that we will be able to identify issues that the programs cannot. The information that I have provided in previous sections indicates that we know quite a lot about where and when automated alignment procedures fail, and thus we know what sorts of problems to look for. This means that it should be possible to develop (and describe) repeatable procedures for manual re-alignment of sequences. Each procedure may be unique to a particular dataset, depending on the characteristics of the starting alignment and the data. These characteristics include the following, all of which can be clearly described and thus meet condition (a) above. These problems are not mutually exclusive, of course.

(1) Inconsistent sequence features. Progressive alignment procedures, in particular, make alignment decisions sequentially, and therefore do not always make these decisions in a consistent manner. For example, highly similar sequence pairs are sometimes misaligned. Also, in protein-coding sequences start and stop codons are often not aligned, and the codon reading frame is sometimes not maintained (especially if there are sequencing errors).

(2) Conserved sequence features that are not aligned. There are many situations where sequences share a similar region but are otherwise not similar, and if there is information about the location of these regions then their alignment should be checked. These regions include: (i) functional sites, such as catalytic regions, sites of intermolecular interactions, substrate binding sites and transcription-factor binding sites; and (ii) structural regions, such as helices, sheets and disulphide bridges in protein-coding sequences, and helices, tetraloops and bulges in RNA-coding sequences. Most conserved

regions are associated with motifs, which can make them easy to locate, but sometimes they involve isolated residues (e.g. cysteines in disulfide bridges).

(3) Single-event sequence blocks that are misaligned. Almost all alignment programs treat each aligned position independently, and thus they cannot be expected to deal correctly with inversions, translocations, transpositions and tandem (or other) duplications. These features create some of the most commonly encountered problems in multiple alignments, especially as most indels are associated with tandem repeats; and so I will discuss them separately below.

(4) Inconsistent decisions. Many alignment decisions are apparently arbitrary, in the sense that several equally good solutions exist irrespective of what optimality criterion is used. Any such arbitrary decisions should be made in a consistent manner. For example, if there are several codons with one nucleotide missing from each, and there is no information that allows a non-arbitrary decision (such as similarity to related sequences), then the gap should be placed in the same position for each codon.

The basic problem that we are trying to avoid by checking for inconsistent alignments is that sometimes the 'phylogenetic informativeness' of an aligned position is a by-product of an incorrect alignment (i.e. informative positions are created where they do not exist). Moreover, these principles apply at all spatial scales. For example, the plant chloroplast genome usually has two copies of an inverted repeat (IR_A, IR_B), and it would be inconsistent if these two regions were not aligned in an identical manner whenever possible. In exchange for avoiding inconsistency that leads to spurious informativeness we also don't want to introduce new but artefactual informativeness, which is the usual downside of any subjective procedure. This can probably be best done by performing any manual re-alignment 'blind' (i.e. without knowledge of the sequence identifications).

As a specific example of possible manual re-alignment of a conserved sequence feature [i.e. problem type (2)], Fig. 1*h* shows part of an amino-acid alignment of the metallo-β-lactamase protein domain-superfamily. At the beginning of the alignment there is a pgHtp motif (where capitalisation indicates an increased degree of conservation) that is known to contain a zinc-binding residue and three active-site residues (see Carfi *et al*. 1995). In the current release of the Pfam database (version 20.0; Finn *et al*. 2006), these sequences are databased in the 'seed' alignment for Lactamase B (accession number PF00753). This alignment contains 324 sequences, and the motif can be located unambiguously in 187 of these sequences (and more ambiguously in many others). However, this functional motif is aligned very erratically among the sequences, varying by up to 11 positions in the alignment. It would be straightforward to manually re-align this motif in an objective and repeatable manner wherever it can be located unambiguously.

As a specific example of possible manual re-alignment of an inconsistent sequence feature [i.e. problem type (1)], the small-subunit rRNA sequences of *Sarcocystis buffalonis* and *Sarcocystis hirsuta* (Apicomplexa) in the alignment of the European rRNA Database (Wuyts *et al*. 2004) have 99.6% nucleotide identity, and yet they are not aligned against each other in three different places: stems E21-3, E21-6 and 47, using the numbering system of Gagnon *et al*. (1996). These misalignments are easy to correct.

The main objective of manual re-alignment is to consider possible mechanistic explanations for the origin of each gap, and then to align the residues so that they reflect the simplest (i.e. most parsimonious) explanation. Step 1 is to identify those characteristics of the gap that might indicate a single origin (e.g. duplication, inversion), and if they exist then use them to guide the alignment. If there are no such characteristics, then Step 2 is to consider whether the sequences should be aligned as a single unit, and either to align them based on similarity, or to unalign them if there is no evidence of homology. In all cases, if there is no evidence to choose among equally plausible alternatives, then you should consistently follow some explicitly stated convention. That is, in the absence of any logical basis for a decision (such as evaluation of empirical evidence) the only way to make the decision repeatable is to adopt an explicitly stated convention.

This suggested procedure for manual re-assessment of a multiple alignment can be best illustrated using an example. Figure 1*A* of Kawakita *et al*. (2003) shows the gapped regions of the arginine kinase (*ArgK*) intron of several *Bombus* species (bumble bees). The authors describe their alignment strategy as: 'we used ClustalX version 1.81 with the default parameter settings. The alignments were then corrected manually for obvious misalignments … [there were] relatively long gaps that were easily aligned … ' Here, I restrict my commentary to those 12 gaps (numbered according to Kawakita *et al*. 2003) involving the 17 ingroup species, as the three outgroup species have several sequence segments that are apparently unrelated to those of any of the ingroup:

Gap-1—for two sequences, a three-base insertion in a well conserved region, so that the location is unambiguous.

Gap-2—for two sequences, deletion of one copy of a four-base perfect tandem repeat (TATT); by convention (see below) the remaining copy has been placed to the left of the gap.

Gap-4—for three sequences, a nine-base deletion in an almost perfectly conserved region, so that the location is unambiguous.

Gap-6—for two sequences, deletion of one copy of an imperfect eight-base tandem repeat (AACTATAA); unfortunately, the remaining copy has been split, so that the initial AA is aligned against the first copy in the other sequences and the CTATAA is aligned against the second copy; this should be corrected for consistency; in this case it is the first copy of the repeat that is imperfect (AACTATAA

in six sequences, AATTATAA in nine sequences and AACCATAA in the outgroup) and so the single copy should be aligned against the perfect copy, to the right of the gap.

Gap-8—for three sequences, a three-base insertion in a perfectly conserved region, so that the location is unambiguous.

Gap-9—for two sequences, a one-base addition to a poly-A repeat; by convention the residues have been left-aligned so that the gap is at the right.

Gap-10—for most of the sequences an AATTA motif is repeated with 63–75 residues in between, but for three of the sequences both of the motifs plus the intervening sequence have been deleted, and replaced with a repeat of the AGT trinucleotide that precedes the first motif in all of the sequences; it is unlikely that the repeat of the AGT is homologous with either of the motif copies and so it should not be aligned against any of the other sequences (i.e. a staggered alignment).

Gap-11—for two sequences, a nine-base deletion in a well conserved region; unfortunately, the gap has been misplaced, so that an ACTATTA motif in these two sequences is aligned against GCTATAA in all of the other sequences instead of their ACTATTA (nine sequences plus outgroup), ACTGTTA (two sequences) or GCTATTA (one sequence) motifs; this should be corrected, thus moving the gap seven positions to the left.

Gap-12—for two sequences, a three-base tandem repeat (ATT); by convention the extra copy has been placed to the right (i.e. the gap in the other sequences is at the right).

Gap-13—for two sequences, an eight-base deletion; by convention the residues have been left-aligned so that the gap is at the right.

Gap-14—for two sequences, a one-base deletion in a poly-A repeat; by convention the residues have been left-aligned; however, this aligns the AA– against AAA (12 sequences), ATA (two sequences) and AAG (one sequence), which is not ideal because it is arbitrary.

Gap-15—for nine sequences, a one-base deletion in a TT dinucleotide; by convention the residues have been left-aligned so that the gap is at the right.

This type of explicit re-assessment of a multiple alignment is both objective and repeatable. It only remains for an author to be clear about how their 'by eye' assessment was carried out—unspecified 'manual adjustments' cannot be considered to be scientific. Some example descriptions of unambiguous procedures from the botanical literature are provided by Kelchner and Clark (1997), Hoot and Douglas (1998) and Graham *et al*. (2000). It is perhaps worth noting that molecular mechanisms creating gaps are likely to be more easily detected in non-coding DNA (Kelchner 2000), such as in my example; e.g. tandem repeats are usually more numerous in non-coding regions. Moreover, a gap does not necessarily have to be created by a single mechanism. For example, in the data of Kreitman (1983) there is a 37-base insertion in three of the 11 sequences, where 19 bp are a direct repeat of the 5′ adjacent sequence and 18 bp are the reverse complement (i.e. an inversion) of a sequence near the 3′ end of the gap—presumably two different evolutionary events are responsible for this pattern.

If objective manual re-alignment of sequences needs to be based solely on sequence similarity (i.e. our mechanistic explanations cannot be used as guides), then what we need is an editor that interactively displays a score for the alignment and each position in it, so that the effect of each manual change can be quantitatively evaluated. Unfortunately, such programs are currently few and far between. There are several programs that display position scores but do not allow manual alignment (such as ClustalX and MACAW); and most sequence editors will colour-code the sequences to show residue conservation or a consensus sequence, which helps visualise the effect of manual re-alignment, but this is not the same as having a quantitative score. Of the several dozen alignment editors that exist, only the following seem to be useful for quantifying the effects of manual re-alignment on similarity: BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), DNAMAN (http://www.lynnon.com/), GeneDoc (http://www.psc.edu/biomed/genedoc/gdpaf.htm), Jalview (http://www.jalview.org/), PAT (http://www-ab.informatik.uni-tuebingen.de/software/pat/welcome.html), PFAAT (http://pfaat.sourceforge.net/) and (perhaps the most adaptable) SQUINT (http://www.cebl.auckland.ac.nz/index.php?target=software&item=6) (all URLs verified 31 October 2006).

The actual criteria that could be used to evaluate manual-alignment quality include the standard optimality scores such as similarity, sum-of-pairs, entropy and log-expectation, all of which would be straightforward to calculate interactively for each position in an editor. Alternatively, we could calculate a single global score for the alignment, with the intention that this global score should increase with each manual change. This is not as practical, however, as it usually involves a complex calculation using a separate program (Thompson *et al*. 2001; Lassmann and Sonnhammer 2005), and there is not yet any consensus about how to do this (Batzoglou 2005). It is also possible to use the quality score of the ensuing phylogenetic tree as the criterion, as is done in the direct optimisation and statistical alignment procedures. The topic of assessing alignment quality is a complex one, which I will not go into here—it needs someone to undertake a proper assessment of the state of the art first.

As a specific example of manual re-alignment of inconsistent decisions [i.e. problem type (4)] based solely on sequence similarity, we can look at the dataset of Rice *et al*. (1997), which is a version of the well known 'Chase *et al.*' dataset, containing 500 aligned sequences of the chloroplast ribulose bisphosphate carboxylase large subunit (*rbcL*) gene from seed plants. Of the characteristics listed above, there are two that seem to be most problematic for these data. First, some of the nucleotides next to (or even in) gapped regions are aligned against a non-conserved nucleotide when there is an adjacent conserved nucleotide of the same type, where conservation is quantified as occurring in ≥95% of the sequences. There are at least 34 such instances that could be re-aligned. The potential problem with these nucleotides is that the original alignment creates parsimony-informative positions that are simply unnecessary artefacts of the alignment itself, whereas the re-alignment reduces the effect of this problem.

Second, some of the gaps are split across two amino acids unnecessarily. This means that two amino acids are 'lost' in the alignment instead of one (i.e. the codons are incomplete), which is an important consideration if the gaps are sequencing artefacts (which they probably are; see below). In at least 41 cases, the nucleotides at the ends of the gap can be re-aligned so that instead there is one 'lost' amino acid plus one conserved amino acid, where conservation is quantified as occurring in ≥95% of the sequences. There are several other instances where re-aligning the gap would create one 'lost' amino acid plus one non-conserved amino acid, but there seems to be no objective justification for making such a re-alignment, as this would create parsimony-informative positions that are unnecessary artefacts of the alignment itself.

If these $34 + 41 = 75$ manual changes are made then we can quantitatively compare the two alignments, to assess objectively whether the new alignment is an improvement over the original one. The new alignment makes very little difference to the average pairwise similarity of the DNA sequences (90.63 *v.* 90.64%) but slightly improves the average pairwise similarity of the amino acid sequences (94.79 *v.* 94.36%). However, the new alignment increases the unweighted parsimony score of the resulting phylogenetic trees based on the nucleotide sequences (9 trees of length 16532 *v.* 16 trees of length 16531, including uninformative characters), which might be considered a retrograde step, depending on whether we focus on the increased score or the reduced number of equally optimal trees. Alternatively, the new alignment improves the weighted parsimony score of the resulting phylogenetic tree based on the amino acid sequences (length 47040 *v.* 47119, based on a step matrix derived from the BLOSUM62 score matrix).

So, whether or not there is a justification for the changes made using these two manual re-alignment criteria can only be decided by first specifying an objective criterion for comparison. The point is, however, that the procedures can both be described in enough detail for them to be repeatable, and their effect can be quantitatively assessed based on specifiable criteria. Manual re-alignment does not *need* to be subjective.

Look for repeats and other sequence blocks

Among the most commonly encountered misalignments in multiple sequence alignments are those caused by single-event sequence blocks, such as inversions, translocations, transpositions and tandem (or other) duplications (Kelchner and Wendel 1996; Benson 1999; Graham *et al*. 2000; Chang and Benner 2004). These features are not unexpected in sequences, as they result from known molecular mechanisms such as slippage during DNA replication / repair and deletion of loop regions in DNA secondary structure (for small blocks), as well as chromosomal processes such as recombination, gene conversion and horizontal gene transfer (for large blocks). However, these features violate the assumptions on which most alignment programs are based, and they can only be dealt with effectively by specialist alignment programs, such as RAlign (Sammeth and Heringa 2006), ABA (Raphael *et al*. 2004) and CombAlign (Wegner *et al*. 2004), unless manual intervention is applied. Hence, these features need to be identified before most automated alignment procedures, and an explicit decision made as to how to deal with them.

Sequence blocks have advantages and disadvantages in a multiple sequence alignment. Within a single sequence, repeats and inversions have the advantage that they can form recognisable anchors in the sequences, which can then be used as constraints to construct an alignment. However, variability between sequences in a multiple alignment can create problems. First, variable numbers of blocks may make homology unclear (e.g. Which block in this sequence should be aligned with which block in the other sequences?). Second, inversions, translocations and transpositions will change the linear order of the sequences (e.g. How does one align a sequence block in one sequence with its reverse complement in another sequence?).

There are thus two practical problems encountered with these sequence blocks. First, the programs treat each block as arising from a series of evolutionary events rather than a single event, and so they do not align the blocks as single units. For example, when there is an unequal number of units (e.g. two repeats in one sequence and three in another) the smaller group of units is usually split across the larger group (e.g. one of the blocks will be split so that it is partly aligned against each of two other blocks, as in one of the examples above). Second, some of these blocks cannot easily be represented by the usual row / column form of alignment. For example, an inversion in one sequence cannot be aligned against an uninverted copy of the same block because the residues will be in a different order. Thus, an alternative representation is needed, such as a cyclic or acyclic graph (Grasso and Lee 2004; Raphael *et al*. 2004; Wegner *et al*. 2004).

Both of these problems need manual attention, but they can be dealt with in an objective and repeatable manner. All that is needed is a specification of how the blocks were detected and a description of how they were dealt with. Tandem repeats are the most common problem encountered in the sorts of sequences used for phylogenetic analysis (along with minute inversions; Kelchner and Wendel 1996), and so I will concentrate on them here.

Microsatellite and minisatellite repeats are usually easy to detect by eye (although there are programs to do it for you: Castelo *et al*. 2002; Anwar and Khan 2006), but longer repeats and non-tandem repeats are more difficult. There are several programs designed to locate longer repeats in a single sequence, such as those of Benson (1999), Kurtz and Schleiermacher (1999), Heger and Holm (2000),

Szklarczyk and Heringa (2004), Campagna *et al*. (2005), Karaca *et al*. (2005), Wexler *et al*. (2005), Achaz *et al.* (2006) and Boeva *et al*. (2006), as well as other sequence re-arrangements (e.g. Darling *et al*. 2004). Having located such elements in a sequence, you then have to check whether they create problems for the between-sequence comparisons (e.g. Is there a variable number of tandem repeats between sequences?). There seems to be little automated help available for such comparisons (although see the VNTRfinder program at http://www.bioinformatics.rcsi.ie/vntrfinder/ and the ProDA program at http://proda.stanford.edu/; both URLs verified 31 October 2006).

Having located the problematic blocks, they need to be dealt with in some manner, keeping in mind that for our purposes the goal is to produce plausible (and parsimonious) hypotheses of potential homology. In the absence of evidence to the contrary, such blocks can be dealt with either by deletion or by convention. The deletion approach tries to find the maximal consistent subset of the sequence lengths that can be aligned by the standard procedures, while the convention approach chooses a (possibly arbitrary) way of re-arranging the sequence blocks so that they can be aligned by the standard procedures. Inversions, for example, could be deleted as literally unalignable (in the usual sense), or a convention could be used that inverts some of the copies so that all of them are in the same orientation and thus alignable (cf. Graham *et al*. 2000). Another convention is to align tandem repeats against the left-hand side of the gapped region, presumably based on the idea that the sequences are transcribed from left to right and the second copy will therefore usually be the 'repeat'. However, this is only a convention. For example, the dataset of Kreitman (1983) has an imperfect 37-nucleotide repeat in some of the sequences, and because the two copies are not identical it is obvious that the first copy is the repeated one (or alternatively the one that has been deleted in some of the sequences).

A more detailed example is shown in Fig. 12, consisting of a set of microsatellite repeats at the tail end of the *Hsp70* gene from various isolates of the gastrointestinal parasite *Cryptosporidium* (Apicomplexa). Here, variations on a GGT GGT ATG CCA motif are repeated eight or more times. This is precisely the sort of situation that conventional computer programs have trouble with, especially when they occur near the beginning or end of an alignment. The first alignment shown in the figure is based on the nucleotide sequences, while the second one is based on first translating the sequences to amino acids and then aligning them. The first alignment has several split amino acids, as well as a split repeat, while both the first and second alignments misalign some obvious later repeated patterns. The third alignment shown is a manual attempt to correct these errors, as well as fixing an obvious slip in the reading frame (sequence AY120919). However, the end of the AF221542 sequence is still of dubious alignment, as are some of the broken repeats (sequences AY120918, AF221538, AF221541). The third alignment is thus the most plausible one as far as evolutionary history is concerned, but it still has improbabilities in it. If these improbabilities cannot be resolved then these positions should presumably be deleted before a tree-building analysis.

Translate to amino acids

It has long been recognised that for biological macromolecules the structure and function are usually more conserved during evolutionary history than is the primary sequence. For example, the biochemical constraints of selective pressures are often assumed to act at the amino-acid level (Fitch and Smith 1983). Consequently, reconstructing evolutionary events may be easier if this secondary and tertiary information is used in addition to the primary sequence information. This point can be seen clearly in Fig. 4, where amino acid alignment is shown to be more reliable then nucleotide alignment as sequence similarity decreases. As a result, it is often suggested that protein-coding sequences should be aligned after translating them to amino acids (and then back-translating the alignment to the equivalent nucleotide sequence for tree building, if desired). This is an objective and repeatable way of increasing the biological insight used in sequence alignment.

However, the actual evolutionary events happen to the DNA, and so there is a lot more information in the nucleotide sequence than in the amino acid sequence of the encoded protein for closely related sequences: in terms of sequence conservation, DNA sequence < protein sequence < protein structure. This is because almost all amino acids are associated with multiple codons, and so there can be nucleotide variation without amino-acid variation. Indeed, serine can be coded without nucleotide alignment conservation at any of the three codon positions, and leucine requires conservation at only the second codon position (the other amino acids require conservation at both the first and second positions). As a typical example, the alignment for the mitochondrial cytochrome c oxidase subunit 1 (*cox1*) gene provided by Cooper *et al*. (2001) for seven ratites and three outgroup birds has 1548 aligned nucleotide positions. In the middle, there is a stretch of 143 amino acids that are conserved across all 10 taxa. However, only 284 of the 429 nucleotide positions are conserved, leaving 145 (33.8%) variable positions. There is thus no phylogenetic information (about the tree topology) at the amino-acid level but there is potentially considerable information at the codon and nucleotide levels. It is for this reason that phylogeny reconstruction from amino acids is not often recommended.

However, it is also for this reason that sequence alignment of protein-coding genes is recommended to be conducted at the amino acid level rather than at the nucleotide level: the alignment is often much easier because of the reduced

Nucleotide alignment

```
                  10        20        30        40        50
AF221529   GGMPGGMPGGMPGGMPGGMPGGMPGGMP?--------------------------
AY120920   GGMPGGMPGGMPGGMPGGMPGGMPGGMPG----------------------------
AY120916   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG----------------------
AF221539   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG?----------------------
AY120918   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGG?---------------
AF221540   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG--
AF221530   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSN--------
AF221533   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNG-----------
AF221536   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG?---------
AF221532   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMSGSNG?------
AF221531   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMSGS?------------
AF221534   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSN----
AF221538   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMGGSN------------
AY120917   GGMPGGMPGG----MPGGMPGGMPGGMPGGMPGGMPGGMPGSN?-----------
AF221537   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP?IYM?--------------
AF221535   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNGPTG
AF221528   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNGPTVEEVD
AF247536   GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP??LMVQL?-----------
AY120919   GGMPGGMPGGMPGGMPGGMPGGMPGG?PGGLPGGMP??LM?--------------
AF221541   GA??GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGPTVEEV---
AF221542   GA??GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMSEGPAIGI?-------
AF221543   GA??GGMPGG-----------------------------------------------
```

Amino acid alignment

```
                  10        20        30        40        50
AF221529   G---GMPGGMPGGMPGGMPGGMPGGMPGGMP?--------------------------
AY120920   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPG----------------------
AY120916   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG----------------------
AF221539   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG?----------------------
AY120918   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGG?---------------
AF221540   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG--
AF221530   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSN--------
AF221533   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNG-----------
AF221536   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG?---------
AF221532   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMSG----SNG?--
AF221531   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMSGS?------------
AF221534   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSN----
AF221538   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMGGSN------------
AY120917   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSN?----------
AF221537   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSIW--------------
AF221535   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNGPTG
AF221528   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNGPTVEEVD
AF247536   G---GMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGSNGPTV-----------
AY120919   G---GMPGGMPGGMPGGMPGGMPGGMPGGCQEVCQEVCQDLM?---------------
AF221541   GAAGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGPTVEEV-------
AF221542   GAAGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMSEGPAIGI?----------
AF221543   GAAGGMPGG-----------------------------------------------
```

Alignment of repeats

```
                  10        20        30        40        50        60
AF221529   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMP?---------------------------------
AY120920   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPG-------------------------------
AY120916   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG-----------------------------
AF221539   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG?-----------------------------
AY120918   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG---GG?-------------------
AF221540   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG----------
AF221530   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP--------GSN--------
AF221533   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP------------GSNG-------
AF221536   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP--------G?---------
AF221532   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMS--------GSNG?-------
AF221531   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMS------------GS?--------
AF221534   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP----GSN--------
AF221538   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGM-G-----------GSN--------
AY120917   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP----------------GSN?-------
AF221537   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP----------------GSIW-------
AF221535   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP----GSNGPTG----
AF221528   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP--------GSNGPTVEEVD
AF247536   ---GGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMP----------------GSNGPTV----
AY120919   ---GGMPGGMPGGMPGGMPGGMPGGMPGG?PGGLPGGMP----------------GSN?-------
AF221541   GAAGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPG--------------GPTVEEV-
AF221542   GAAGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGMPGGM------------------SEGPAIGI?-
AF221543   GAAGGMPGG-----------------------------------------------------------
```

Fig. 12. Three alternative amino-acid alignments of the end of the 70-kDa heat shock protein (*Hsp70*) gene from 22 isolates of *Cryptosporidium* (Apicomplexa). This demonstrates the problems that sequence repeats (paralogues) can cause for automated alignment procedures. The first alignment is based directly on the nucleotides, and the second alignment is based on first translating the nucleotides to amino acids (both calculated by ClustalW 1.83 with default settings). The third alignment is a manual re-arrangement of the second alignment, emphasising the pattern and variation in the repeat nature of the sequences. A '?' indicates an unknown amino acid due to either a deleted or a non-sequenced nucleotide. The original nucleotide data are from Xiao *et al.* (2002).

variability (i.e. it can be easier to find plausible locations for indels). The other main reasons for using amino acids are the increased character-state space (i.e. 20 amino acids *v.* 4 nucleotides; Simmons *et al*. 2004), and that the gaps will be inserted in groups of three nucleotides, thus maintaining the reading frame of the sequences (which single-nucleotide indels will destroy). For the same reasons, amino acids are also recommended for database searching, even if the original data are nucleotide sequences (Wernersson and Pedersen 2003).

As an example to illustrate the potential advantages for sequence alignment, Fig. 13 shows part of the DNA polymerase A gene from various species of the endoparasite *Leishmania* (Kinetoplastida). The first alignment is based on the nucleotide sequences, while the second one is based on first translating the sequences to amino acids and then aligning them, followed by back-translation. The nucleotide alignment tries to minimise the number of gaps in the first four sequences as well as to maximise the similarity of the first triplet after the gap (i.e. it aligns CAC against GAC rather than against CGC / CGT / CGA). However, CGC / CGT / CGA all code for arginine, a basic amino acid, while GAC codes for aspartic acid, an acidic amino acid. Since CAC codes for histidine, another basic amino acid, the amino acid alignment maximises the similarity in the type of amino acid. The second alignment seems to be more biologically plausible, since it hypothesises evolutionary events that will have less effect on the structure of the resulting protein.

This re-alignment of the first gap creates a second indel in the affected sequences, which is most parsimoniously aligned with the second indel in the other three sequences. This second indel is created by a lack of repeat somewhere in a set of GAC / GAT (aspartic acid) repeats or a set of GAG / GAA (glutamic acid) repeats. The amino acid alignment places the gap at the boundary between the two series of repeats, as this maximises the consistency of the first and last columns of the repeats (i.e. the gaps are placed in a mixture of Asp and

Nucleotide alignment

```
                  1       10        20        30        40        50        60
L. aethiopica     AAGCGGGGC-----------------CACGACGACGAGGAGGAGGATGGCAAACGGAAG
L. arabica        AAGCGGGGC-----------------CACGACGACGAGGAGGAGGATGGCAAACGGAAG
L. tropica        AAGCGGGGC-----------------CACGACGACGAGGAGGAGGATGGCAAACGGAAG
L. major          AAGCGGGGC-----------------CACGACGACGAGGAGGAGGATGGCAAACGGAAG
L. chagasi        AAGCGGGGCCGC--------------GACGACGAGGAGGAGGAGGATGGGAAGCGGAAG
L. infantum       AAGCGGGGCCGT--------------GACGACGAGGAGGAGGAGGATGGGAAGCGGAAG
L. donovani       AAGCGGGGCCGC--------------GACGACGACGAGGAGGAGGATGGGAAGCGGAAG
L. mexicana       AAGCGGGGCCGC--------------GACGACGACGAGGAGGAGGATGGCAAGCGGAAG
L. amazonensis    AAGCGGGGCCGC--------------GACGACGACGAGGAG---GATGGCAAGCGGAAG
L. gymnodactyli   AGGCGGGGTCGA--------------GACGACGATGAAGAGGAGGATGGCAAGCGGAGA
L. tarentolae     AGGCGGGGTCGA--------------GACGACGATGAAGAGGAGGATGGCAAGCGGAGA
L. hoogstraali    AGGCGGGGTCGA--------------GACGACGATGAAGAGGAGGATGGCAAGCGGAGA
L. adleri         AGGCGGGGTCGA--------------GACGACGACGAAGAGGAGGATGGCAAGCGGAGG
L. braziliensis   AAGCGGGGCCGC--------------GACGACGACGAGGAG---GATGGCAAGCGGAAG
L. panamensis     AAGCGGGGCCGC--------------GACGACGACGAGGAG---GATGGCAAGCGGAAG
L. deanei         AAGCGCAGCCGC--------------GACGATGAGGGAGAGGAGGATGGCAAGCGGAAG
L. hertigi        AAGCGCAGCCGC--------------GACGATGAGGGAGAGGAGGATGGCAAGCGGAAG
L. herreri        AAGCGCAGCCGTGGCGGTGGCGGCGACGAGGATGACGGGGAGGAAGATGGCAAACGAAAG
E. monterogeii    AAGCGCAGCCGTGGCGGTGGCGGCGACGACGACGACGGGGAGGAAGATGGCAAGCGAAAG
```

Amino acid alignment

```
                  1       10        20        30        40        50        60
L. aethiopica     AAGCGGGGCCAC--------------GACGAC---GAGGAGGAGGATGGCAAACGGAAG
L. arabica        AAGCGGGGCCAC--------------GACGAC---GAGGAGGAGGATGGCAAACGGAAG
L. tropica        AAGCGGGGCCAC--------------GACGAC---GAGGAGGAGGATGGCAAACGGAAG
L. major          AAGCGGGGCCAC--------------GACGAC---GAGGAGGAGGATGGCAAACGGAAG
L. chagasi        AAGCGGGGCCGC--------------GACGACGAGGAGGAGGAGGATGGGAAGCGGAAG
L. infantum       AAGCGGGGCCGT--------------GACGACGAGGAGGAGGAGGATGGGAAGCGGAAG
L. donovani       AAGCGGGGCCGC--------------GACGACGACGAGGAGGAGGATGGGAAGCGGAAG
L. mexicana       AAGCGGGGCCGC--------------GACGACGACGAGGAGGAGGATGGCAAGCGGAAG
L. amazonensis    AAGCGGGGCCGC--------------GACGAC---GACGAGGAGGATGGCAAGCGGAAG
L. gymnodactyli   AGGCGGGGTCGA--------------GACGACGATGAAGAGGAGGATGGCAAGCGGAGA
L. tarentolae     AGGCGGGGTCGA--------------GACGACGATGAAGAGGAGGATGGCAAGCGGAGA
L. hoogstraali    AGGCGGGGTCGA--------------GACGACGATGAAGAGGAGGATGGCAAGCGGAGA
L. adleri         AGGCGGGGTCGA--------------GACGACGACGAAGAGGAGGATGGCAAGCGGAGG
L. braziliensis   AAGCGGGGCCGC--------------GACGAC---GACGAGGAGGATGGCAAGCGGAAG
L. panamensis     AAGCGGGGCCGC--------------GACGAC---GACGAGGAGGATGGCAAGCGGAAG
L. deanei         AAGCGCAGCCGC--------------GACGATGAGGGAGAGGAGGATGGCAAGCGGAAG
L. hertigi        AAGCGCAGCCGC--------------GACGATGAGGGAGAGGAGGATGGCAAGCGGAAG
L. herreri        AAGCGCAGCCGTGGCGGTGGCGGCGACGAGGATGACGGGGAGGAAGATGGCAAACGAAAG
E. monterogeii    AAGCGCAGCCGTGGCGGTGGCGGCGACGACGACGACGGGGAGGAAGATGGCAAGCGAAAG
```

Fig. 13. Two alternative alignments of part of the DNA polymerase A gene from 18 *Leishmania* and an *Endotrypanum* species (Kinetoplastida). This illustrates the advantages of aligning protein-coding sequences as amino acids rather than as nucleotides. The first alignment is based directly on the nucleotides, and the second alignment is based on first translating the nucleotides to amino acids (both calculated by ClustalW 1.83 with default settings). The original nucleotide data are from Croan *et al.* (1997).

Glu amino acids). The nucleotide alignment places the gaps at the end of the second series of repeats. Since the sequences are 'read' from left to right, it seems more plausible to place such gaps at the end of the series of repeats, although there is no reason why repeats could not be added or subtracted at any position.

Thus, there are distinct advantages to aligning amino acids rather than nucleotides. However, the major problem with using amino acids to align nucleotides is simply that it assumes that the *only* cause of gaps in the alignment is insertion or deletion of an amino acid. Clearly, there are two other serious possibilities.

(1) Sequencing artefacts or curation artefacts (such as misplaced introns, intron / exon boundaries or start and stop signals). A good illustrative example is the dataset of Rice *et al*. (1997) described above. For the 500 aligned sequences, there are 423 internal gaps (as well as many terminal ones) among the 1398 aligned nucleotide positions, 406 of which are shown in the graph of Fig. 14 (the remaining gaps are 15–180 nucleotides long). The frequency distribution fits a logarithmic series extremely well, implying that a single process has created this set of gaps. Clearly, given the gap lengths, insertion or deletion of amino acids is not that process, since these would create gaps in multiples of three nucleotides. Indeed, it seems more likely to be the quality of the sequencing that is the culprit here, as the gaps are concentrated in relatively few (96 of 500) of the sequences, and the data date from 1993 when laboratory techniques were less sophisticated than they are now. (The number of in-frame stop codons is also a bit of a give-away.)

(2) A true slip in the reading frame that is quickly corrected and thus affects only a few amino acids. Two simple examples are shown in Fig. 15, based on adding two outgroup taxa to some of the *Cryptosporidium* data discussed above. The first example (the actin gene) is apparently quite straightforward from the point of view of the protein sequences (the top alignment in the figure), as there is a single amino-acid indel (probably an insertion) in the *Plasmodium falciparum* II sequence, followed by an amino-acid substitution one position later. However, from the point of view of the DNA sequences these two differences involve a series of changes over the nine positions involved (positions 16–24); even the unchanged amino acid in the middle involves a nucleotide substitution. Therefore, it is more parsimonious to postulate a quite different evolutionary history for the DNA sequences (the bottom alignment), involving re-alignment of all of the other sequences relative to *P. falciparum* II. This replaces all of the substitutions with a smaller number of indel events. The consequence for constructing a phylogenetic tree is that the pattern of shared character states among the sequences is very different between the two alignments, and so they support different trees.

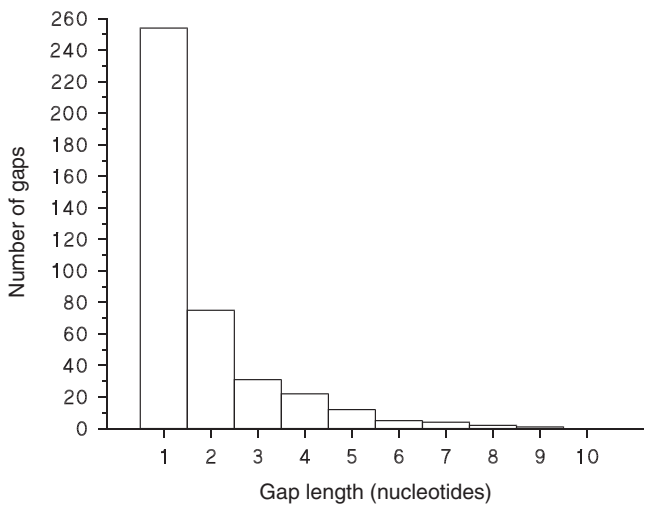

Fig. 14. Frequency histogram of the gap size for 406 of the 423 internal gaps in the 500-taxon dataset of Rice *et al.* (1997). The histogram fits a logarithmic series extremely well, implying that a single process has created this set of gaps. The remaining gaps are 15–180 nucleotides long and do not fit the series, suggesting that they were created by one or more additional processes.

The second example (the *Hsp70* gene) is in some ways more subtle, because there is no length variation in the amino-acid sequences, but it has an equally big effect. From the protein perspective there are two amino-acid substitutions (and no indels) in the *P. falciparum* sequence compared to the others. However, from the DNA perspective these changes involve five nucleotide substitutions out of six positions (top alignment). Therefore it is more parsimonious to postulate an insertion–deletion pair in the DNA sequence, which causes a short frame-shift, and no substitutions at all (bottom alignment). These are obviously quite different hypotheses of the molecular evolution. However, there is also a practical importance for phylogenetic tree building, because in the bottom alignment all of the character-state differences are unique to *P. falciparum*, and so will probably not affect the choice of tree topology, while in the top alignment position 13 has a character state shared between *P. falciparum* and *Cryptosporidium canis*, which can quite definitely affect the choice of tree.

In both of these examples translating the sequences to amino acids to perform the alignment is probably counter-productive, even if there is apparently no length variation in the unaligned sequences. In fact, from this point of view it is actually quite difficult to work out what the word 'homology' means with reference to an amino acid sequence when frame-shifts have occurred. Note that the homology hypotheses in

```
Amino acid alignment          Actin                          Hsp70
                    1       10        20        30      1       10        20
P. falciparum_II    AAACGTTCTGAAGAACATTCAGATGAAATAGAA
P. falciparum_I     AAAACATCTGAACAA---AGCAGTGATATTGAA   TTAGATGTT-TGCTCCTTATCATTA
B. bovis            AACTCTTCTGCATCA---TCTAGTGAAATCGAG   CTCGATGTC-GCTCCACTCTCCCTC
C. parvum_human     AAGAAATCTCAAGAA---TCTTCTGAATTAGAG   TTGGATGTT-GCTCCATTATCACTC
C. canis_dog        AAGAAGTCTCAGGAG---TCTTCAGAATTAGAA   TTGGATGTT-GCCCCACTGTCCCTC
C. baileyi          AAGAAATCACAGGAA---TCATCTGAACTTGAA   TTAGATGTT-GCTCCATTATCACTT
C. felis            AAAAAGTCTCAGGAG---TCTTCCGAACTTGAA   CTGGATGTT-GCTCCTTTGTCTCTC
C. wrairi           AAGAAATCTCAAGAA---TCTTCTGAATTAGAG   TTGGATGTT-GCTCCATTATCACTC
C. saurophilum      AAGAAATCTCAAGAA---TCTTCTGAGCTTGAA   TTGGATGTT-GCTCCTTTGTCTCTT
C. meleagridis      AAGAAATCTCAAGAA---TCTTCTGAATTAGAG   TTGGATGTT-GCTCCATTATCACTT

Nucleotide alignment          Actin                          Hsp70
                    1       10        20        30      1       10        20
P. falciparum_II    AAACGTTCTGAAGAACATTCAGATGAAATAGAA
P. falciparum_I     AAAACATCTGAACAAAG--CAG-TGATATTGAA   TTAGATGTTTGCTCC-TTATCATTA
B. bovis            AACTCTTCTGCATCATCT--AG-TGAAATCGAG   CTCGATGTC-GCTCCACTCTCCCTC
C. parvum_human     AAGAAATCTCAAGAATCTTC---TGAATTAGAG   TTGGATGTT-GCTCCATTATCACTC
C. canis_dog        AAGAAGTCTCAGGAGTCTTC---AGAATTAGAA   TTGGATGTT-GCCCCACTGTCCCTC
C. baileyi          AAGAAATCACAGGAATCATC---TGAACTTGAA   TTAGATGTT-GCTCCATTATCACTT
C. felis            AAAAAGTCTCAGGAGTCTTC---CGAACTTGAA   CTGGATGTT-GCTCCTTTGTCTCTC
C. wrairi           AAGAAATCTCAAGAATCTTC---TGAATTAGAG   TTGGATGTT-GCTCCATTATCACTC
C. saurophilum      AAGAAATCTCAAGAATCTTC---TGAGCTTGAA   TTGGATGTT-GCTCCTTTGTCTCTT
C. meleagridis      AAGAAATCTCAAGAATCTTC---TGAATTAGAG   TTGGATGTT-GCTCCATTATCACTT
```

Fig. 15. Partial alignments of the actin and 70-kDa heat-shock protein (*Hsp70*) genes for seven species of *Cryptosporidium* plus *Babesia bovis* and *Plasmodium falciparum* (Apicomplexa). This illustrates the disadvantages of aligning protein-coding sequences as amino acids rather than as nucleotides. There are two actin genes for *P. falciparum* (paralogues I and II) but only one gene for *Hsp70*. The alignments were derived using the ClustalW 1.83 program, with default settings. Two alternative alignments are shown for each gene. The bottom alignment was created using the DNA sequence data, while the top alignment was created after the DNA data had been translated to the equivalent amino-acid sequence (and then back-translated to nucleotide sequences). The original nucleotide data are from Xiao *et al.* (2002).

the amino acid alignment are wrong for *Hsp70*, because the cysteine (TGC) is not homologous to the alanines (GCT, GCC) and nor is the serine (TCC) homologous with the prolines (CCA, CCT). At the nucleotide level, the serine T is homologous to the alanine T while the serine CC is homologous to the proline CC, and so the serine has no simple homologue among the other amino acids (and neither does the cysteine).

So, if you choose to align nucleotides as translated amino acids then you should always check the nucleotide sequences afterwards, to see if the potential homologies are plausible and parsimonious. Note, also, that in possibility (2) the gap represents an indel, while in (1) it does not. This emphasises the point that gaps and indels are not the same thing.

There are several computer programs that have been designed to automate the translation, alignment and back-translation of protein-coding sequences. These include MRTrans (ftp.virginia.edu in pub/fasta/other/mrtrans.shar; verified 2 November 2006), RevTrans (Wernersson and Pedersen 2003), CodonAlign (http://www.sinauer.com/hall; verified 2 November 2006), and transAlign (Bininda-Emonds 2005). Only the latter program attempts to deal explicitly with shifts in the reading frame, which it does not always do successfully. Aligning protein-coding nucleotide sequences in a manner that is robust to frame-shifts is not easy, even for pairwise comparisons (Arvestad 1997).

There is also the matter of aligning nucleotide sequences that only partially code for proteins. That is, we have long had the ambition to align sequences of coding and non-coding DNA, with or without frame-shifts, and possibly with multiple reading frames (Hein 1994). Several usable algorithms for pairwise alignment under these circumstances have been developed (Hein and Støvlbæk 1996; Pedersen *et al*. 1998; Hua *et al*. 1999), but the only attempt to produce multiple alignments is that of Stocsits *et al*. (2005), with the CodAln program.

Structure-based alignment

In an earlier section, I argued that it is possible to apply to molecular sequences the same principles and practice of detailed structural and functional (e.g. biochemical, biophysical and genetic) analyses of characters that have traditionally been used for the assessment of homology in studies of phenotypic attributes (these studies being based on the observed close relationship between structure and function in biology). More to the point, I contend that it is now quite easy to do so in practice. It is thus usually unnecessary for phenetic pattern-matching procedures to continue to dominate molecular phylogeny, because the sequence–structure–function relationship provides a straightforward mechanism for explicitly incorporating evolutionary homology into molecular sequence alignment. This circumstance does not yet seem to have been widely appreciated by practicing phylogeneticists, because the

structure-based multiple-sequence alignments have mainly been used for predicting the molecular structures themselves, rather than for phylogenetic analysis.

A deeper understanding of the secondary and tertiary structures of the molecules should contribute to a better understanding of evolutionary homology, in the same way that studies of ontogeny and developmental constraints have contributed to the assessment of primary homology for phenotypic data. However, similarity of topology and shared functional constraints only represent *evidence* on which to base hypotheses of homology, and they may still actually represent homoplasy instead. The argument for using structural considerations is based on the idea that it is unlikely that a group of structurally related genes would arise independently (Gough 2005), and so they are very likely to have evolved from a common ancestor. However, the hypothesised homologues may result from functional convergence (i.e. analogy), and this may not necessarily be unlikely for the residues of structural genes (e.g. see the example discussed in detail by Sadreyev and Grishin 2004).

Nevertheless, for RNA-coding sequences at least, the use of putative secondary-structure models is very likely to produce multiple-sequence alignments that are close to the true alignments, in the sense of having aligned evolutionarily homologous nucleotides (Kjer 1995; Hickson *et al*. 1996; Morrison and Ellis 1997), since the higher-order structures inferred from comparative analyses are now quite refined. This will be especially important in situations where the phylogenetic signal is weak (Taylor 1986; Kjer 1995), such as in remotely related taxa, since small errors in the data matrix can then have large effects on the subsequent phylogenetic inferences.

In particular, the existence of plausible structure models provides an explicit and repeatable criterion for aligning variable regions of the sequences (Taylor 1986; Kelchner 2000). These regions have traditionally been considered to have ambiguous alignments, because there can be multiple equally optimal alignments in these areas if maximising similarity is the only criterion used. Furthermore, these ambiguously aligned regions are often excised from phylogenetic studies because of their presumed unreliability. (Note that this is different from simply excluding gapped regions from subsequent analyses, since gapped regions are not necessarily ambiguously aligned.) However, if the structure models can be successfully used for alignment then the ambiguities are reduced (along with the multiple optima), because the acceptable alignment is chosen only from among those that maintain the conserved structural features; for example, compensatory base changes in rRNA structures provides evidence on which to base hypotheses of homology. Under these circumstances, exclusion of portions of the sequences could then be based on their apparent lack of phylogenetic information rather than on artefacts such as pattern-matching ambiguities. Finally, the structure models also provide a valuable framework for checking the accuracy of the individual sequences themselves (i.e. proofreading), since the consistency of conserved motifs must be maintained (Taylor 1986; Kjer 1995, 2004; Hickson *et al*. 1996; Gillespie *et al*. 2005*b*).

A specific example is shown in Fig. 16, based on the sequences of the α and β chains of human haemoglobin. These two genes are considered to be the product of an ancient duplication and thus are paralogous. The proteins also fold into almost identical structures. However, the alignment as shown has some ambiguities in it, which cannot easily be resolved by either nucleotide or amino-acid similarity. These can be dealt with by making the alignment reflect the structural and functional relationships between the two chains. It is principally the α chain that requires gaps in order to align it to the longer β chain, and so it is the gaps in the α chain that need adjusting. The DLS amino-acid motif near position 150 appears to be well aligned between the two chains, but the side-chain of the histidine (H) near position 160 of the α haemoglobin occupies approximately the same structural space as that of the methionine (M) at position 173 of the β chain. More importantly, these two side-chains have the same function, which is to 'glue' this part of the amino-acid sequence to the core of the protein (which the histidine does by a salt bridge and the methionine does via hydrophobic contacts). So, both structural and functional considerations suggest aligning the H with the M, which are currently shown at opposite ends of the gap. Note that this contradicts the alignments presented by both Fitch and Smith (1983) and Knudsen and Miyamoto (2003).

As a more extreme example, we can consider aligning the large-subunit rRNA genes from the mitochondrial genomes of the 13 species that have currently been sequenced from nematodes. The 11 sequences from the class Chromadorea have 61–84% pairwise nucleotide identity, which means that they are difficult to align accurately using sequence similarity alone. However, they are all easy to align (even manually) against the sequence alignment (for three of the species) and secondary structure (for one species) provided by the CRW database (Cannone *et al*. 2002). On the other hand, the two sequences from the class Enoplea have only 40–48% pairwise nucleotide identity with the other sequences (sharing only a couple of small motifs), and they are thus very difficult to align even under these circumstances. Nevertheless, they can both be aligned with the other sequences based on their own secondary-structure diagrams (provided in the original publications), as all of the structures are still very similar. This can be done either manually or using a computer program that aligns RNA secondary structures. It is, of course, another matter whether a phylogenetic analysis based on primary sequence patterns would be productive at this low level of

```
                1       10        20        30        40        50        60        70
               ATGGTG---CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTAT
Human_alpha     M  V  -  L  S  P  A  D  K  T  N  V  K  A  A  W  G  K  V  G  A  H  A  G  E  Y
Human_beta      M  V  H  L  T  P  E  E  K  S  A  V  T  A  L  W  G  K  V  -  -  N  V  D  E  V
               ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTG------AACGTGGATGAAGTT

                80        90       100       110       120       130       140       150
               GGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTC---GACCTGAGC
Human_alpha     G  A  E  A  L  E  R  M  F  L  S  F  P  T  T  K  T  Y  F  P  H  F  -  D  L  S
Human_beta      G  G  E  A  L  G  R  L  L  V  V  Y  P  W  T  Q  R  F  F  E  S  F  G  D  L  S
               GGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCC

                160       170       180       190       200       210       220       230
               CAC---------------GGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTG
Human_alpha     H  -  -  -  -  -  G  S  A  Q  V  K  G  H  G  K  K  V  A  D  A  L  T  N  A  V
Human_beta      T  P  D  A  V  M  G  N  P  K  V  K  A  H  G  K  K  V  L  G  A  F  S  D  G  L
               ACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTG

                   240       250       260       270       280       290       300       310
               GCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTC
Human_alpha     A  H  V  D  D  M  P  N  A  L  S  A  L  S  D  L  H  A  H  K  L  R  V  D  P  V
Human_beta      A  H  L  D  N  L  K  G  T  F  A  T  L  S  E  L  H  C  D  K  L  H  V  D  P  E
               GCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAG

                     320       330       340       350       360       370       380       390
               AACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCAC
Human_alpha     N  F  K  L  L  S  H  C  L  L  V  T  L  A  A  H  L  P  A  E  F  T  P  A  V  H
Human_beta      N  F  R  L  L  G  N  V  L  V  C  V  L  A  H  H  F  G  K  E  F  T  P  P  V  Q
               AACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAG

                       400       410       420       430       440       450
               GCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAA
Human_alpha     A  S  L  D  K  F  L  A  S  V  S  T  V  L  T  S  K  Y  R  *
Human_beta      A  A  Y  Q  K  V  V  A  G  V  A  N  A  L  A  H  K  Y  H  *
               GCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAA
```

Fig. 16. Alignment of the human haemoglobin α and β chains, showing both the DNA and amino acid sequences. The alignment was produced by ClustalW version 1.83 based on the amino acid sequences, and does not reflect the structural and functional relationships between the two chains. The example is modified from Rodriguez and Vriend (1997).

identity. It is just as likely that the conservation of structure is a result of functional constraints instead of evolutionary homology (see Schultes *et al*. 1999), in which case the sequence alignment will be of little use for phylogenetic purposes.

It also needs to be recognised that the alignment of sequences against structure models for proteins and RNAs is not perfect, and the placement of some residues will remain arbitrary. For example, the helices and strands are the structurally conserved regions of proteins while the loops and coils are more loosely conserved, and consequently alignment of the helical and strand regions is much more straightforward (Poch and Delarue 1996). Alternatively, for RNAs it is the loops and bulges that are usually well conserved (because of their role in protein recognition), while the double-stranded parts of the helices are usually the easiest to align because of the compensatory base changes needed to maintain the stems (Varani and Pardi 1994). As an example, in Fig. 16 it seems that at the nucleotide level the two valines (V) at position 5 are aligned, but in the folded protein the α-chain valine actually occupies the same structural position as the adjacent histidine (H) of the β chain. Here, secondary structure and primary similarity contradict each other unresolvably.

There can also be potential problems of non-homology in the alignments, which can result, for example, from variability in the size and position of helices, local structural rearrangements caused by nearby mutations, mutation 'hotspots' where reversals and parallelisms are common, phenomena such as replication slippage, or simply misapplication of the method due to subjective bias (Kjer 1995; Hickson *et al*. 1996; Hancock and Vogler 2000; Kelchner 2000; Gillespie 2004). An obvious example is the hypervariable regions that occur in rRNA, where there is extreme length variation (e.g. Gillespie *et al*. 2005*b*), and replication slippage often leads to convergence on similar primary and secondary structures (see the instances discussed by Hancock and Vogler 2000 and Shull *et al*. 2001). Homology assessment in such regions can be difficult or impossible, and such regions may be best left unaligned. There is also non-homology due to functional constraints, as discussed above.

Thus, it is important to emphasise here that I am not advocating the use of pure structure alignments as phylogenetic alignments. Structure multiple alignments are useful for structure prediction but not necessarily for phylogenetic analysis. Instead, what I am referring to is structural consistency, which can be thought of as a structure-based alignment. Although there is no simple definition of a structure-based alignment (e.g. what components it must or must not have), the distinction can be made clear with

reference to an alignment of RNA sequences, as shown in Fig. 17. Here, three possible alignments are shown, with the first one based solely on sequence similarity and the second one based solely on structural similarity (these are thus optimal alignments based on two different criteria). In the similarity alignment, the hypotheses of homology are complex, as three of the aligned residues in each sequence are postulated to have changed their functional role in the formation of the short stem. This is a rather unlikely scenario, and it is certainly not parsimonious. In the structure alignment, equivalent structural residues are aligned, so that three gap positions are introduced into each sequence in order to align those residues that are paired in the stem (i.e. the complementary base pairs). This scenario thus postulates at least four indels, which may not be any more likely evolutionarily, and it is not much more parsimonious. The structurally consistent alignment, on the other hand, bases the alignment on sequence similarity with the proviso that if one half of a stem-pair is aligned then the other half must also be aligned. It thus differs from the structure alignment in not insisting that all paired residues be aligned. In this case, only one indel is postulated in each sequence, both involving a change in function for the other residue, plus one other change in function in each sequence.

A specific example of a structure-based alignment is shown in Fig. 18, for 22 *ITS2* rRNA gene sequences from two families of mites. This form of alignment presentation makes it clear where the homology problems are and what they are, when trying to align sequences across taxonomic groups. Thus the plausibility of hypotheses of homology can be evaluated; and unalignable regions can be explicitly defined. The visual presentation of such alignments for RNA sequences is discussed in more detail by Kjer *et al*. (1994), Kjer (1995) and Gillespie (2004).

In this example, it is usually considered that *ITS2* molecules form a four-stem structure, as discussed by Schultz *et al*. (2005), with the stems labelled I–IV. The whole structure forms a closed loop, because the preceding (5′) sequence in the 5.8S molecule pairs with the succeeding (3′) sequence in the 28S molecule. The *ITS2* is thus easily excised from the growing ribosomal macromolecule by simply cutting one stem.

In rRNA sequences, many of the helices are easily aligned both within and between taxonomic groups, as they are often length-invariant and very similar due to the functional necessity of base pairing; even the single-stranded bulges can usually be aligned. For example, in this example helices 0 and III are unproblematic (Fig. 18). However, helices I and IV are difficult to locate, with the I′ and IV′ sides apparently straightforward to align but not the opposite side of these stems.

Some of the characteristic features of the *ITS2* also occur only sporadically in these sequences. For example, the unpaired U–U in helix II exists in all Phytoseiidae species except for *Metaseiulus occidentalis* but not in the Rhinonyssidae (where there is an A–T or G–T pair). The same taxonomic distribution occurs for the UGGU motif at the base of helix III.

Moreover, in this example helix II–II′ is easily aligned for the Phytoseiidae but not for the Rhinonyssidae, and thus it is difficult to align across both taxonomic groups. Also, helices can be regions of expansion or contraction (REC), where parts of the helix are not homologous between groups, such as helix II–II′ in this example, where the paired region IIa–IIa′ is common to both families but region IIb–IIb′ is unique to the Phytoseiidae. Helices can also help to determine which alignment gaps are likely to be sequencing artefacts, since each helix base must pair with another base. For example, the four-base gap in helix 0′ of *Neoseiulus californicus* must be an artefact, as the corresponding paired nucleotides exist in helix 0.

The helix terminal (hairpin) loops of rRNA are usually easily aligned within taxonomic groups but are often doubtful between groups, even when they are length-invariant. For example, the terminal loop of helix II–II′ is six bases long in both families but they are doubtfully homologous. The initial TG of most of the Rhinonyssidae sequences may actually be homologous with the initial unpaired TG between helices IIa and IIb of the Phytoseiidae; and, indeed, the similarity-based sequence-alignment programs all align this TG motif across the two groups.

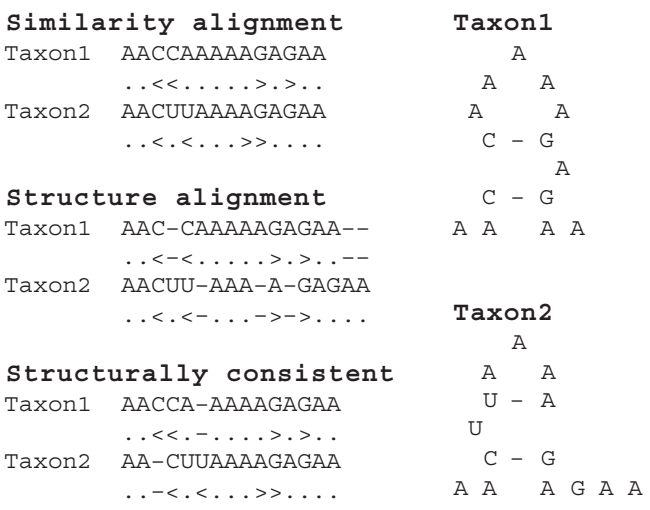

Fig. 17. An artificial example of the difference between pairwise alignment of two RNA sequences based on similarity, structure and structural consistency. On the left, each of the two sequences is shown along with its secondary structure, where complementary paired residues in a short stem are indicated by angle brackets, so that '<' indicates the 5′ residue of the pair and '>' indicates the 3′ residue. A schematic diagram of each stem structure is shown on the right. Each alignment proposes a different set of hypotheses concerning the homology of the residues between the two stems. This example is based on one published by Gardner and Giegerich (2004).

	0 (5.8S)		I		I′		IIa		IIb		IIb′		IIa′
Ascoidea: Phytoseiidae													
Iphiseius degenerans	CATTTGTTTCAG(T)ATA	TAAAC--	CGA	[ATC--AT]	ACG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	A-CGCTAT	GCAA(T)GGTA
Typhlodromus pyri	CATTTGTTTCAG(T)ATA	TAAAC--	CGT	[ATC--AT]	AAG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	ATCGCTAT	GCAA(T)GGTA
Phytoseiulus persimilis	CATTTGTTTCAG(T)ATA	TAAAC--	CGT	[ATC--AT]	AAG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	ATCGCTAT	GCAA(T)GGTA
Euseius finlandicus 1	CATTTGTTTCAG(T)ATA	TAAAC--	CGA	[ATC--AT]	AAG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	AACGCTAT	GCAA(T)GGTA
Euseius finlandicus 2	CATTTGTTTCAG(T)ATA	TAAAC--	CGA	[ATC--AT]	AAG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	AACGCTAT	GCAA(T)GGTA
Euseius concordis	CATTTGTTTCAG(T)ATA	TAAAC--	CGA	[ATC--AT]	AAG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	A-CGCTAT	GCAA(T)GGTA
Neoseiulus cucumeris	CATTTGTTTCAG(T)ATA	TAAAC--	CGT	[ATC--AT]	ACG	TATT	TACC(T)TTGC	TGCA	GCC	[CTCGTC]	GGC	ACCGCTAT	GCAA(T)GGTA
Neoseiulus fallacis	CATTTGTTTCAG(T)ATA	TAAAC--	CGT	[ATC--AT]	ACG	TATT	TACC(T)TTGC	TGCA	GCC	[CTTGTC]	GGC	AACGCTAT	GCAA(T)GGTA
Neoseiulus californicus	CATTTGTTTCAG(T)ATA	TAAAC--	CGT	[ATC--AT]	ACG	TATT	TACC(T)TTGC	TGCA	GCC	[CTTGTC]	GGC	AACGCTAT	GCAA(T)GGTA
Metaseiulus occidentalis	CATTTGTTTCAG(T)ATA	TAATC--	CGT	[ATC--AT]	ACG	TATT	TACT(G)TTGC	TGCG	GCC	[CTCGCT]	GGC	GAAGCTAT	GCAA(T)AGTA
Dermanyssoidea: Rhinonyssidae													
Tinaminyssus sartbaevi	CACTTGTTTCAG(T)ATA	TTAAA--	CGA	[AGCA-AT]	AAG	TAAT	TACT(A)TTGC	----------	[TGCAAC]	----------			GCAA(T)GGTA
Tinaminyssus streptopelioides	CATTTGTTTCAG(T)ATA	AAAAA--	GTA	[CTA--AT]	GCG	TAGT	TACT(A)TTGC	----------	[TGCAAC]	----------			GCAA(T)AGTA
Tinaminyssus bubulci	CATTTGTTTCAG(T)ATA	CAAA---	CAA	[GCA--AT]	ACG	TAAT	TACT(A)TTGC	----------	[TGAAAC]	----------			GCAT(T)AGTA
Tinaminyssus melloi	CATTTGTTTCAG(T)ATA	CAAAA--	CTA	[GCA--AA]	AAG	TAGT	TACT(A)TTGC	----------	[TACAAC]	----------			GCAT(T)AGTA
Tinaminyssus minisetosum	CACTTGTTTCAG(T)ATA	TTGAC--	CTA	[GCA--TT]	ACG	TAGT	TACT(A)TTGC	----------	[TGCAAT]	----------			GCAA(T)AGTA
Tinaminyssus columbae	CACTTGTTTCAG(T)ATA	TTGA---	CAT	[AGAA-AT]	ACG	TAGT	TACT(A)TTGC	----------	[TGCGAT]	----------			GCAA(T)AGTA
Sternostoma strandtmanni	CATTTGTTTCAG(T)ATA	TAAAAGA	TGT	[CCA--AT]	ACG	TAAT	TGCT(G)TTGC	----------	[TGGAAA]	----------			GCAA(T)GGCA
Sternostoma boydi	CATTTGTTTCAG(T)ATA	TAAAAGA	TGT	[CCA--AT]	ACG	TAAT	TGCT(G)TTGC	----------	[TGGAAA]	----------			GCAA(T)GGCA
Sternostoma fulicae	CATTTGTTTCAG(T)ATA	TATAAAA	TGT	[TCA--AT]	ACG	TACT	TGCT(G)TTGC	----------	[TGGGAG]	----------			GCAA(T)GGTA
Tropilaelaps koenigerum	CACTTGTTTCAG(T)ATA	TAACT--	CGT	[CGT--AT]	AAG	TACT	GACT(A)TTGC	----------	[CGTTAT]	----------			GCAA(T)GGTA
Tropilaelaps clareae	CACTTGTTTCAG(T)ATA	TAACT--	CGT	[AGT--AT]	ATG	TACT	TACT(A)TTGC	----------	[CG-TAC]	----------			GCAA(T)GGTA
Dermanyssus gallinae	CATTTGTTTCAG(T)ATA	TAT----	CAA	[TGTAAAT]	TTG	TCGT	TACT(A)TTGC	----------	[TTGAAA]	----------			GCAA(T)AGTA

		III		III′		IV		IV′		0′ (28S)
Ascoidea: Phytoseiidae										
Iphiseius degenerans	TAAATTCTCATT	GGTCAC	[GAGA]	GTGA(TA)CC	AACC---	-CA	[AACCTTTA]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Typhlodromus pyri	TAAATTCTCTTT	GGTCAC	[AAGA]	GTGA(TA)CC	AAAC---	-CG	[AACCAATA]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Phytoseiulus persimilis	TAAATTCTCTTT	GGTCAC	[AAGA]	GTGA(TA)CC	AAAC---	-CG	[AACCAATA]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Euseius finlandicus 1	TAAATTCTCATT	GGTCAC	[GAGA]	GTGA(TA)CC	AACC---	-CA	[AACCATTA]	T??	????	?????????(?)??????
Euseius finlandicus 2	TAAATTCTCATT	GGTCAC	[GAGA]	GTGA(TA)CC	AACC---	-CA	[AACCATTA]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Euseius concordis	TAAATTCTCATT	GGTCAC	[GAGA]	GTGA(TA)CC	AACC---	-CA	[A-CC-TTA]	TGA	CGTG	TATCTGAAA(?)??????
Neoseiulus cucumeris	TAAATTCTCTTT	GGTCAC	[AAGA]	GTGA(TA)CC	AAAC---	-CA	[AACCATTA]	TGA	CGTG	TATTTGAAA(T)CAAGTG
Neoseiulus fallacis	TAAATTTTCTTT	GGTCAC	[AAGA]	GTGA(TA)CC	AAAC---	-CA	[AACCTTTA]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Neoseiulus californicus	TAAATTTTCTTT	GGTCAC	[AAGA]	GTGA(TA)CC	AAAC---	-CA	[AACCTTTA]	TGA	CGTG	TATC----A(T)CAAGTG
Metaseiulus occidentalis	TAAATTCTCAAT	ATTCAC	[GTGA]	GTGA(TA)??	????---	-??	[????????]	???	????	?????????(?)??????
Dermanyssoidea: Rhinonyssidae										
Tinaminyssus sartbaevi	GAAATATTCTAT	GATCAC	[AAGA]	GTGA(TA)TT	GTGCTT-	GAG	[AGT-----]	GAG	TGAG	TATCTGAAA(T)CAAGTG
Tinaminyssus streptopelioides	GAAGTTACGTTT	AATCAC	[AAGA]	GTGG(TG)TC	AACAGA-	ATG	[AGT-----]	TAA	CGTG	TATCTGAAA(T)CAAGTG
Tinaminyssus bubulci	GAAGTTTTCACG	GATCAC	[AGGA]	GTGG(TG)TA	ACAAAAT	TTA	[ATGATT--]	TAA	CGTG	TATCTGAAA(T)CAAGTG
Tinaminyssus melloi	GAAATTTTTCAT	GATCAC	[AAGA]	GTGA(TA)AC	AAGCGA-	TGG	[AGT-----]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Tinaminyssus minisetosum	TAAGTTTTCCAT	AATCAC	[AAGA]	GTGA(TG)CT	AATCTG-	TGT	[AAT-----]	GGA	CGTG	TATCTGAAC(T)CAAGTG
Tinaminyssus columbae	TAAGTTTTCCTT	AATCAG	[AAGT]	GTGA(TG)CT	AATCCG-	TGT	[AAT-----]	GGA	CGTG	TATCTGAAC(T)CAAGTG
Sternostoma strandtmanni	CAAGATCTCTTG	AATCAC	[AAGA]	GTGA(CA)TT	AGTCAA-	TTG	[GTG-----]	TAG	CGTG	TATCTGAAA(T)CAAGTG
Sternostoma boydi	CTAGATCTCTTG	AATCAC	[AAGA]	GTGA(CA)TT	AGTCAA-	TTG	[GTG-----]	TAG	CGTG	TATCTGAAA(T)CAAGTG
Sternostoma fulicae	TAAGATCTATTG	AATCAC	[AAGA]	GTGA(CG)TT	AGTCAG-	TTG	[ATG-----]	TAG	TGTG	TATCTGAAA(T)CAAGTG
Tropilaelaps koenigerum	TAAAATCTCGAC	GGTCAC	[GAGA]	GTGA(TG)CT	GCCTGC-	TCA	[AGT-----]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Tropilaelaps clareae	TAAAATCTCCAC	GGTCAC	[GAGA]	GTGA(TG)CT	GCCTGC-	TCA	[AGT-----]	TGA	CGTG	TATCTGAAA(T)CAAGTG
Dermanyssus gallinae	TAAACTGTTCGC	GATCAC	[AAGT]	GTGA(TA)TC	GATCTT-	GCG	[TTATA---]	TGA	CGTG	TATCTGAAA(T)CAAGTG

Fig. 18. DNA sequences of the internal transcribed spacer 2 (*ITS2*) gene of rRNA plus the flanking parts of the 5.8S and 28S genes from several species of two families of mites (Mesostigmata: Dermanyssina). The sequences have been aligned to illustrate the structural features of the folded RNA. Here, '| |' delimit paired regions, as labelled (e.g. the region labelled 0, which is part of the 5.8S, pairs with region 0′, which is part of the 28S); '()' represent bulges in helices; and '[ ]' demarcate the terminal (hairpin) loops of helices. The sequences are aligned within these marked regions but not necessarily aligned outside of them. Note that in some helices there are non-canonical base-pairs, but that there are also compensatory substitutions. A '?' indicates an unknown nucleotide.

The other unpaired regions (i.e. the connecting sequences between helices) sometimes cannot even be aligned easily within taxonomic groups. In these cases the sequences are shown unaligned, the nucleotides simply being moved to one end or the other of the region (called regions of ambiguous alignment, RAA). However, in this example the regions between helices I′ and IIa, IIa′ and III, and IV′ and 0′ are length-invariant and very similar across both groups, and thus are shown aligned here.

The objective of a structure-based alignment is thus to delimit the various structural regions of the end product molecule (e.g. protein or RNA), which defines a set of sequence zones within each of which homology can be assessed separately. Some of these zones will have clear primary-sequence similarity and some will not. Among the latter zones will be those that can be aligned even in the absence of sequence similarity, as well as those that cannot. The aligned zones can be used as anchor points irrespective of whether there is any sequence conservation or not, followed by further attempts at alignment using primary sequence similarity, if desired.

Incorporating structure information into alignment

In order to incorporate structural information into multiple-alignment protocols, as suggested above, there are three possible strategies: (i) perform a manual alignment according to some structure model; (ii) incorporate structure-based parameters into an automated alignment procedure; and (iii) base the alignment directly on a pre-existing structure-based alignment. The latter two avenues have recently been explored in the literature, with the first one being better explored for protein-coding sequences and the latter for RNA-coding sequences.

For proteins, many experimentally determined (e.g. by NMR) three-dimensional structural models now exist, and are available in the Protein Data Bank (Westbrook *et al*. 2003). It is recognised that the organisation of these molecules into helices, strands and loops, coils or turns results in unequal probabilities of indels occurring in each of these three functional regions (Pascarella and Argos 1992), and alignment procedures for protein-coding genes should take this phenomenon into account [i.e. strategy (ii)]. Computerised alignment algorithms have been developed to do this, by having position-specific gap penalties (which control the number of indels inferred in the sequences) rather than having an average value that is applied throughout the sequence (e.g. Henneke 1989; Bell *et al*. 1993; Higgins *et al*. 1996; Taylor 1996), or by including the amino acid properties (such as hydrophobicity, polarity, size, charge) as weights (e.g. Taylor 1986; Johnson *et al*. 1990; Zhang and Kahveci 2006). It is also possible to use information from structure-prediction programs directly in an alignment strategy (Al-Lazikani *et al*. 2001; Jennings *et al*. 2001; O'Sullivan *et al*. 2004; Simossis and Heringa 2005; Zhou and Zhou 2005; Armougom *et al*. 2006).

For RNA-encoding sequences, models of structure and function are now very well developed indeed (Higgs 2000), both experimentally determined for smaller molecules and derived from comparative sequence analysis (also called homology modelling) for larger molecules (Gutell *et al*. 2002). The models provide details of the secondary-structure arrangement of the RNA molecules into helices (with double-stranded stems as well as single-stranded bulges and loops) and non-helical regions; and the tertiary structure of some of the molecules is also now beginning to be addressed (Gutell *et al*. 2002). There are many programs available to predict RNA secondary structure (see Gardner and Giegerich 2004), and some of these simultaneously predict structure and produce a multiple alignment (e.g. Hofacker *et al*. 2004; Touzet and Perriquet 2004; Bauer *et al*. 2005*a*, 2005*b*; Holmes 2005; Siebert and Backofen 2005; Bafna *et al*. 2006; Dalli *et al*. 2006), although this is a very difficult process, especially for long genes.

Alternatively, strategy (i) suggests that these structure models for proteins and RNAs could be used to aid manual alignment of the sequences. For example, Ponting and Birney (2000) give some advice for manually improving a multiple alignment in relation to conserved structural features of proteins. However, in practice manual alignments have been restricted almost entirely to RNA-coding sequences (other than manual alignment of protein-coding sequences that are almost identical). This effectively involves comparative sequence analysis of new sequences against pre-existing structure models, searching for compensatory base changes (Kjer *et al*. 1994). Kjer (1995) gives details of a manual method for structure-based RNA alignment (see also the jRNA website: http://hymenoptera.tamu.edu/rna/index.php; verified 31 October 2006; Gillespie *et al*. 2005*a*); and suitable templates can be developed based on sequence motifs (Hickson *et al*. 1996; Kjer 1997; Kelchner 2002; Gillespie *et al*. 2005*b*). This method is detailed and explicit enough so that empirical evidence indicates it to be at least as repeatable as any automated procedure (Kjer *et al*. 2006).

More important, though, is strategy (iii), which points out that multiple-sequence alignment can be performed by automatically aligning new sequences against a database of sequences that have themselves already been aligned against known structures. That is, the database sequences are used as alignment templates (or seed alignments) to guide subsequent alignment procedures, thus explicitly using the structural models to improve the homology of the resulting alignments (Rost and Valencia 1996). This can be thought of as 'jump-starting' alignment (cf. Mecham *et al*. 2006), where all of the hard work done to produce

previous alignments is not wasted but is instead used as the starting point for later work. Personally, I (and others, e.g. Nicholas *et al*. 2002) think that this approach will be the way of the future for multiple sequence alignment for phylogenetic purposes (and probably also for phylogenetic analysis in general). To this end, there is a growing trend to provide curated databases of aligned genome sequences for particular groups of organisms that can be used as alignment templates (e.g. HIV Sequence Database: http://hiv-web.lanl.gov/content/index; Database of Homologous Sequences from Complete Genomes: http://pbil.univ-lyon1.fr/databases/hogenom.html; Organellar Genome Retrieval system: http://drake.physics.mcmaster.ca/ogre/; GreenGenes: http://greengenes.lbl.gov/; all URLs verified 31 October 2006).

For protein-coding sequences, perhaps the most useful of the general databases for the creation of multiple alignments is the Pfam database (Finn *et al*. 2006), which has alignments of all of the sequences from the Swiss-Prot / TrEMBL amino acid databases, these alignments being grouped based on the protein structural domains from the ProDom database. If you want an amino acid alignment for a gene then you can simply download an alignment from the database, although the alignment will be for only those parts of the sequence with a recognised (i.e. classified) structural domain. This set of downloaded sequences can be used as a template for aligning your own sequences. Nucleotide alignments of these same sequences are available in the Pandit database (Whelan *et al*. 2006), which may thus be of more direct relevance for phylogenetic analysis.

Similarly, online databases of the known (i.e. published) sequences now exist for many RNAs in the Rfam database (Griffiths-Jones *et al*. 2005), as well as in specialised databases for the 5S (Szymanski *et al*. 2002) and small- and large-subunit (Cannone *et al*. 2002; Cole *et al*. 2005; Wuyts *et al*. 2004; DeSantis *et al*. 2006*a*) rRNA genes, as well as for the tRNA (Helm *et al*. 2000; Rainaldi *et al*. 2003; Sprinzl and Vassilenko 2005), tmRNA (Gueneau de Novoa and Williams 2004), uRNA (Zwieb 1997), SRP-RNA (Andersen *et al*. 2006), and RNase P RNA (Brown 1999) genes. All of these databases store the sequences in a format based on the alignment inferred from the secondary-structure models. For these DNA sequences there is thus no need to use computerised pattern-matching algorithms to produce multiple-sequence alignments—the alignments are simply downloaded from the appropriate database.

There are also non-coding sequences that have characteristic secondary structure. These include: introns, such as the Group II *cis*-splicing introns which have a stem–loop structure that is necessary for autosplicing (Kelchner 2002); spacers, such as the internal transcribed spacers in RNA-coding regions (Schultz *et al*. 2005) and inter-genic spacers (Kelchner 2000); and the many types of non-coding RNA now being recognised (Eddy 2002*b*). Where these sequences are conserved within a taxonomic group, and are presumably therefore under some form of evolutionary constraint, they are useful phylogenetic markers. Unfortunately, these sequences are less likely to have had their structure elucidated for your study group (see Schultz *et al*. 2005), and so you may have to do it yourself (as I did above for the mite *ITS2* sequences). There are currently no specialised databases of alignments available for access, although Rfam has some alignments of intron domains.

Unfortunately, none of the databases that I have listed is yet set up perfectly for phylogenetic analysis. For example, the protein databases are usually arranged by conserved domains, and most proteins have two or more domains. This means that the gene sequences will have somewhat different names in different parts of the database, and sometimes have non-obvious abbreviated names. Also, the sequences may have no direct structure annotation; and the non-conserved sequences between the domains are not included. Moreover, the particular structure models chosen for each database can influence the alignment and therefore the phylogenetic inference (Winnepenninckx and Backeljau 1996; Marchler-Bauer *et al*. 2002).

Perhaps most importantly, none of these databases is perfectly curated, and so manual checking of the aligned sequences for consistency among closely related taxa is still necessary. For example, in the current release of the Rfam database (version 7.0; Griffiths-Jones *et al*. 2005), the 5.8S sequence alignment in accession number RF00002 has a consensus secondary structure diagram (for *Homo sapiens*) that does not match the secondary structure diagram shown in the cited literature reference. Alternatively, the alignments provided by the databases are not necessarily either structurally consistent or consistent with their structure model. For example, there are three *Plasmodium falciparum* (Apicomplexa) sequences in the 5.8S sequence alignment of the Comparative RNA Website (Cannone *et al*. 2002), and these have two stems where there are inconsistencies. In the first stem (sometimes called B7), sequence U48228 has a structure diagram with a five-base hairpin loop and is aligned that way, but sequence U21939 is aligned as a five-base loop but shows a structure with a three-base loop. In the second stem (sometimes called B8), U21939 has a structure and an alignment with a four-base hairpin loop, U48228 has a structure with a four-base loop but is aligned as a six-base loop, and sequence AL031746 is indicated as having a six-base loop that is not aligned with either of the other two sequences. *Caveat emptor*.

Use of these databases requires you to align your new sequences against a pre-existing alignment. Some of

the web services provided with the databases allow you to align a new sequence against their database online, including the Pfam (Finn *et al*. 2006), Rfam (Griffiths-Jones *et al*. 2005), RDP-II (Cole *et al*. 2005), HPV (http://hpv-web.lanl.gov/stdgen/virus/hpv/; verified 31 October 2006), NAST (DeSantis *et al*. 2006*b*) and European rRNA (http://pbil.univ-lyon1.fr/databases/rnali.html; verified 31 October 2006) databases, which can be effective if you have a small number of sequences. These methods often use hidden markov models (or their extensions, called stochastic context-free grammars) rather than profiles, so that the resulting alignments can be of high quality. You can also adopt this strategy yourself by using the programs directly (Eddy 1998, 2002*a*). For protein-coding sequences, it may be sufficient simply to use the profile-alignment option of a standard progressive alignment program (Jennings *et al*. 2001). Alternatively, several computerised alignment algorithms have been developed to align new RNA sequences against a sequence or pre-existing alignment based on both primary and secondary structures (e.g. Corpet and Michot 1994; Notredame *et al*. 1997; Lenhof *et al*. 1998); Page (2000) discusses an alignment web-server based on this idea. Alternatively, one can treat the structure alignment as a profile and then use position-specific gap penalties and sequence weighting to align this to new sequences (e.g. O'Brien *et al*. 1998; Thébault *et al*. 1999).

If either strategy (i) or (ii) is being used, then it is helpful if the multiple sequence alignment editor has the capability to interface with a program that deals with structures. For protein-coding sequences, such editors include AntheProt (Deléage *et al*. 1988), InterAlign (Pible *et al*. 2005), Jalview (Clamp *et al*. 2004), STRAP (Gille and Frömmel 2001) and ViTO (Catherinot and Labesse 2004). For RNA sequences, it is necessary that the editor takes into account and displays the secondary structure, so that structural consistency can be maintained in paired regions. There are a few such programs, including BioEdit (Hall 1999), but there are also specialist editors such as DCSE (De Rijk and De Wachter 1993) and RALEE (Griffiths-Jones 2005). It also helps if you can extract the structure information from the databases along with the sequences (see Telford *et al*. 2005).

If strategy (iii) is being used, then it is helpful to have an appropriate tool to maintain your own databases of alignments, to which you can align your new sequences. The only such tools available at the moment are ARB (Ludwig *et al*. 2004), jPHYDIT (Jeon *et al*. 2005) and RibAlign (Teeling and Gloeckner 2006), with the former being the most generally useful package.

Conclusions

Sequence alignment is often seen as a bioinformatics procedure rather than a biological one. Thus, many previous reviews of multiple alignment have concentrated on descriptions of algorithms rather than on biological principles. I have tried to redress the balance by covering a series of topics that seem to me to have been under-stressed or under-valued, particularly within the context of phylogenetic analysis.

First, sequence alignment in a phylogenetic analysis is about assessing homologies: the residues aligned should be homologous in the evolutionary sense. These homologies are hypotheses about unknowable evolutionary events rather than empirical observations. Thus, a phylogenetic alignment can differ considerably from other forms of sequence alignment. For molecular structure prediction we need to align structurally equivalent residues, and for sequence comparison we need to align functionally equivalent motifs. Both of these criteria are amenable to empirical quantification, and they do not necessarily involve the alignment of residues that share evolutionary descent—analogy can be just as effective as homology for these alignments. For database searching we need to align residues that maximise the difference between homologous and non-homologous sequences, rather than homologous residues. This can be quantified statistically to some extent, and it also does not necessarily involve homology of residues.

Second, homology assessment for sequence alignment should involve information from whatever source is appropriate. Traditionally, sequence similarity has been the primary criterion for assessing residue homology, but this becomes increasingly inadequate as sequence identity decreases. Therefore, information from structural and functional studies needs to be incorporated into sequence alignment procedures, not as a replacement for sequence similarity but as an adjunct to it. Decisions about residue homology can be based on sequence similarity, structural similarity and functional similarity.

Third, much is now known about the circumstances under which computer programs based on progressive sequence alignment will fail. While a lot of this information comes specifically from studies of amino acid sequences rather than nucleotide sequences, it still has direct relevance to phylogenetic studies. The pattern-matching algorithms in the commonly used programs were designed for sequence-comparison alignment rather than for phylogenetic alignment, and so they cannot be expected to produce alignments suitable for reconstructing phylogenies except under specific circumstances. I have provided an original summary of this information, which should be taken into account when planning a phylogenetic analysis. Common problems include terminal and internal gaps, orphan sequences and strong subset groupings, and repeated sequence blocks.

Fourth, there have been many recent developments in the matter of producing multiple alignments, but few

of these have yet had much impact on phylogenetic studies. I have summarised the recent attempts in the sequence-comparison literature to improve progressive sequence alignment procedures, especially for amino acid sequences. Instinctively reaching for your copy of Clustal is not necessarily the wisest choice of alignment tool. I have also summarised recent attempts to synthesise alignment and tree building, thus producing a coherent one-step phylogenetic analysis. There is still some way to go before these become practical procedures, but some enthusiasm from the customers would not go astray.

Fifth, I have provided some explicit suggestions for increasing the biological insight that is employed when constructing a multiple sequence alignment for phylogenetic purposes. Most of these suggestions consist of nothing more than taking into account known evolutionary processes when making alignment decisions, thus supplementing simple sequence similarity with information from a broader context. The most important of these suggestions involves incorporating the sequence–structure–function relationship into the alignment procedure. These procedures can be objective and repeatable, and can involve computerised algorithms to automate much of the work. However, I have not offered any explicit protocols for implementing any of these methods, as each dataset needs to be taken on its own merits. Human quality control should not be ignored in science, and human judgements cannot be avoided when erecting hypotheses.

Finally, alignment should not be seen as a process that is started anew for every dataset. Alignment should be seen as a process where new sequences are added to a pre-existing alignment that has been manually curated by the biologist. That is, the time and effort that has gone into producing alignments of high quality for phylogenetic purposes should be added to rather than discarded, by using previous alignments as templates for subsequent alignments.

References

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

Aboitiz F (1987) Letter to the editor. *Cell* **51**, 515–516. doi: 10.1016/0092-8674(87)90117-6

Achaz G, Boyer F, Rocha EPC, Viari AC (2006) Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics*, (in press). doi: 10.1093/bioinformatics/bt1519

Al-Lazikani B, Sheinerman FB, Honig B (2001) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of janus kinases. *Proceedings of the National Academy of Sciences USA* **98**, 14 796–14 801. doi: 10.1073/pnas.011577898

Allison L, Wallace CS (1994) The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *Journal of Molecular Evolution* **39**, 418–430. doi: 10.1007/BF00160274

Allison L, Wallace CS, Yee CN (1992) Minimum message length encoding, evolutionary trees and multiple alignment. In 'Proceedings of the Hawaii international conference on system sciences (HICSS-25).' pp. 663–674. (IEEE Press: Piscataway)

Althaus E, Caprara A, Lenhof H-P, Reinert K (2002) Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics. *Bioinformatics* **18**, S4–S16.

Anbarasu LA, Narayanasamy P, Sundararajan V (2000) Multiple molecular sequence alignment by island parallel genetic algorithm. *Current Science* **78**, 858–863.

Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, Wower J, Gorodkin J, Samuelsson T, Zwieb C (2006) The tmRDB and SRPDB resources. *Nucleic Acids Research* **34**, D163–D168. doi: 10.1093/nar/gkj142

Anwar T, Khan AU (2006) SSRscanner: a program for reporting distribution and location of simple sequence repeats. *Bioinformation* **1**, 89–91.

Apostolico A, Giancarlo R (1998) Sequence alignment in molecular biology. *Journal of Computational Biology* **5**, 173–196.

Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Research* **34**, W604–W608.

Arvestad L (1997) Aligning coding DNA in the presence of frame-shift errors. *Lecture Notes in Computer Science* **1264**, 180–190.

Badger JH, Eisen JA, Ward NL (2005) Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and Caulobacterales. *International Journal of Systematic and Evolutionary Microbiology* **55**, 1021–1026. doi: 10.1099/ijs.0.63510-0

Bafna V, Tang H, Zhang S (2006) Consensus folding of unaligned RNA sequences revisited. *Journal of Computational Biology* **13**, 283–295. doi: 10.1089/cmb.2006.13.283

Bahr A, Thompson JD, Thierry J-C, Poch O (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research* **29**, 323–326. doi: 10.1093/nar/29.1.323

Barta JR (1997) Investigating phylogenetic relationships within the Apicomplexa using sequence data: the search for homology. *Methods* **13**, 81–88. doi: 10.1006/meth.1997.0501

Barton GJ, Sternberg MJE (1987) A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *Journal of Molecular Biology* **198**, 327–337. doi: 10.1016/0022-2836(87)90316-0

Batzoglou S (2005) The many faces of sequence alignment. *Briefings in Bioinformatics* **6**, 6–22. doi: 10.1093/bib/6.1.6

Bauer M, Klau GW, Reinert K (2005*a*) Fast and accurate structural RNA alignment by progressive lagrangian optimization. *Lecture Notes in Computer Science* **3695**, 217–228.

Bauer M, Klau GW, Reinert K (2005*b*) Multiple structural RNA alignment with lagrangian relaxation. *Lecture Notes in Computer Science* **3692**, 303–314.

Baumel A, Ainouche ML, Bayer RJ, Ainouche AK, Misset MT (2002) Molecular phylogeny of hybridizing species from the genus *Spartina* Schreb. (Poaceae). *Molecular Phylogenetics and Evolution* **22**, 303–314. doi: 10.1006/mpev.2001.1064

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 Molecular Biology* **9**, 515–520. doi: 10.1046/j.1365-2583.2000.00211.x

Bell LH, Coggins JR, Milner-White EJ (1993) Mix'n'Match: an improved multiple sequence alignment procedure for distantly related proteins using secondary structure predictions, designed to be independent of the choice of gap penalty and scoring matrix. *Protein Engineering* **6**, 683–690.

Belshaw R, Quicke DLJ (2002) Robustness of ancestral state estimates: evolution of life history strategy in ichneumonoid parasitoids. *Systematic Biology* **51**, 450–477. doi: 10.1080/10635150290069896

Benner SA, Cohen MA, Gonnet GH (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology* **229**, 1065–1082. doi: 10.1006/jmbi.1993.1105

Benson G (1997) Sequence alignment with tandem duplication. *Journal of Computational Biology* **4**, 351–367.

Benson G (1999) Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580. doi: 10.1093/nar/27.2.573

Bininda-Emonds ORP (2005) TransAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* **6**, 156. doi: 10.1186/1471-2105-6-156

Bishop MJ, Thompson EA (1986) Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology* **190**, 159–165. doi: 10.1016/0022-2836(86)90289-5

Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biology* **6**, 0030.

Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences USA* **83**, 5155–5159. doi: 10.1073/pnas.83.14.5155

Bledsoe AH, Sheldon FH (1990) Molecular homology and DNA hybridization. *Journal of Molecular Evolution* **30**, 425–433. doi: 10.1007/BF02101114

Boeva V, Regnier M, Papatsenko D, Makeev V (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* **22**, 676–684. doi: 10.1093/bioinformatics/btk032

Bonizzoni P, Della Vedova G (2001) The complexity of multiple sequence alignment with SP-score that is a metric. *Theoretical Computer Science* **259**, 63–79. doi: 10.1016/S0304-3975(99)00324-2

Brawley SH (1999) Submission and retrieval of an aligned set of nucleic acid sequences. *Journal of Phycology* **35**, 433–437. doi: 10.1046/j.1529-8817.1999.3520433.x

Brenner SE, Chothia C, Hubbard TJ (1998) Assessing sequence comparison methods with reliable structurally-identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences USA* **95**, 6073–6078. doi: 10.1073/pnas.95.11.6073

Briffeuil P, Baudoux G, Lambert C, De Bolle X, Vinals C, Feytmans E, Depiereux E (1998) Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. *Bioinformatics* **14**, 357–366. doi: 10.1093/bioinformatics/14.4.357

Britten RJ, Rowen L, Williams J, Cameron RA (2003) Majority of divergence between closely related DNA samples is due to indels. *Proceedings of the National Academy of Sciences USA* **100**, 4661–4665. doi: 10.1073/pnas.0330964100

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

Brown JW (1999) The ribonuclease P database. *Nucleic Acids Research* **27**, 314. doi: 10.1093/nar/27.1.314

Bucka-Lassen K, Caprani O, Hein J (1999) Combining many multiple alignments in one improved alignment. *Bioinformatics* **15**, 122–130. doi: 10.1093/bioinformatics/15.2.122

Butler AB, Saidel WM (2000) Defining sameness: historical, biological, and generative homology. *BioEssays* **22**, 846–853. doi: 10.1002/1521-1878(200009)22:9<846::AID-BIES10>3.0.CO;2-R

Campagna D, Romualdi C, Vitulo N, Del Favero M, Lexa M, Cannata N, Valle G (2005) RAP: a new computer program for de novo identification of repeated sequences in whole genomes. *Bioinformatics* **21**, 582–588. doi: 10.1093/bioinformatics/bti039

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, *et al*. (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 2. doi: 10.1186/1471-2105-3-2

Carfi A, Pares S, Duée E, Galleni M, Duez C, Frère JM, Dideberg O (1995) The 3-D structure of a zinc metallo-β-lactamase from *Bacillus cereus* reveals a new type of protein fold. *EMBO Journal* **14**, 4914–4921.

Cartmill M (1994) A critique of homology as a morphological concept. *American Journal of Physical Anthropology* **94**, 115–123. doi: 10.1002/ajpa.1330940109

Cartwright RA (2005) DNA assembly with gaps (DAWG): simulating sequence evolution. *Bioinformatics* **21**, iii31–iii38. doi: 10.1093/bioinformatics/bti1200

Castelo AT, Martins W, Gao GR (2002) TROLL—tandem repeat occurrence locator. *Bioinformatics* **18**, 634–636. doi: 10.1093/bioinformatics/18.4.634

Catherinot V, Labesse G (2004) ViTO: tool for refinement of protein sequence–structure alignments. *Bioinformatics* **20**, 3694–3696. doi: 10.1093/bioinformatics/bth429

Cerchio S, Tucker P (1998) Influence of alignment on the mtDNA phylogeny of Cetacea: questionable support for a Mysticeti / Physeteroidea clade. *Systematic Biology* **47**, 336–344. doi: 10.1080/106351598260941

Chain P, Kurtz S, Ohlebusch E, Slezak T (2003) An applications-focused review of comparative genomics tools: capabilities, limitations, and future challenges. *Briefings in Bioinformatics* **4**, 105–123. doi: 10.1093/bib/4.2.105

Chakrabarti S, Bhardwaj N, Anand PA, Sowdhamini R (2004) Improvement of alignment accuracy utilizing sequentially conserved motifs. *BMC Bioinformatics* **5**, 167. doi: 10.1186/1471-2105-5-167

Chakrabarti S, Lanczycki CJ, Panchenko AR, Przytycka TM, Thiessen PA, Bryant SH (2006) Refining multiple sequence alignments with conserved core regions. *Nucleic Acids Research* **34**, 2598–2606. doi: 10.1093/nar/gkl274

Chan SC, Wong AKC, Chiu DKY (1992) A survey of multiple sequence comparison methods. *Bulletin of Mathematical Biology* **54**, 563–598.

Chang MSS, Benner SA (2004) Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *Journal of Molecular Biology* **341**, 617–631. doi: 10.1016/j.jmb.2004.05.045

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* **31**, 3497–3500. doi: 10.1093/nar/gkg500

Chiaromonte F, Yap VB, Miller W (2002) Scoring pairwise genomic sequence alignments. In 'Proceedings of the 7th Pacific Symposium on Biocomputing 2002, Lihue, Hawaii'. pp. 115–126.

Chindelevitch L, Li Z, Blais E, Blanchette M (2006) On the inference of parsimonious indel evolutionary scenarios. *Journal of Bioinformatics and Computational Biology* **4**, 721–744. doi: 10.1142/S0219720006002168

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview java alignment editor. *Bioinformatics* **20**, 426–427. doi: 10.1093/bioinformatics/btg430

Cognato AI, Vogler AP (2001) Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Systematic Biology* **50**, 758–780. doi: 10.1080/106351501753462803

Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research* **33**, D294–D296. doi: 10.1093/nar/gki038

Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R (2001) Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **409**, 704–707. doi: 10.1038/35055536

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* **16**, 10 881–10 890.

Corpet F, Michot B (1994) RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Computer Applications in the Biosciences* **10**, 389–399.

Cozzetto D, Tramontano A (2005) Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins: Structure, Function, and Bioinformatics* **58**, 151–157. doi: 10.1002/prot.20284

Croan DG, Morrison DA, Ellis JT (1997) Evolution of the genus *Leishmania* revealed by comparison of DNA and RNA polymerase gene sequences. *Molecular and Biochemical Parasitology* **89**, 149–159. doi: 10.1016/S0166-6851(97)00111-4

Dalli D, Wilm A, Mainz I, Steger G (2006) STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics* **22**, 1593–1599. doi: 10.1093/bioinformatics/btl142

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**, 1394–1403. doi: 10.1101/gr.2289704

De Laet JE (2005) Parsimony and the problem of inapplicables in sequence data. In 'Parsimony, phylogeny, and genomics.' (Ed. VA Albert) pp. 81–116. (Oxford University Press: Oxford)

Deléage G, Clerc FF, Roux B, Gautheron DC (1988) ANTHEPROT: a package for protein sequence analysis using a microcomputer. *Computer Applications in the Biosciences* **4**, 351–356.

De Rijk P, De Wachter R (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. *Bioinformatics* **9**, 735–740.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006*a*) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**, 5069–5072. doi: 10.1128/AEM.03006-05

DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL (2006*b*) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research* **34**, W394–W399. doi: 10.1093/nar/gkj156

Dewey CN, Pachter L (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Human Molecular Genetics* **15**, R51–R56. doi: 10.1093/hmg/ddl056

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research* **15**, 330–340. doi: 10.1101/gr.2821705

Domingues FS, Lackner P, Andreeva A, Sippl MJ (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *Journal of Molecular Biology* **297**, 1003–1013. doi: 10.1006/jmbi.2000.3615

Donoghue MJ, Sanderson MJ (1994) Complexity and homology in plants. In 'Homology: the hierarchical basis of comparative biology'. (Ed. BK Hall) pp. 393–421. (Academic Press: San Diego)

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? *Science* **214**, 149–159. doi: 10.1126/science.7280687

Duret L, Abdeddaïm S (2000) Multiple alignments for structural, functional, or phylogenetic analyses of homologous sequences. In 'Bioinformatics: sequence, structure, and databanks.' (Ed. D Higgins, W Taylor) pp. 51–76. (Oxford University Press: Oxford)

Ebedes J, Datta A (2004) Multiple sequence alignment in parallel on a workstation cluster. *Bioinformatics* **20**, 1193–1195. doi: 10.1093/bioinformatics/bth055

Eddy SR (1998) Profile hidden markov models. *Bioinformatics* **14**, 755–763. doi: 10.1093/bioinformatics/14.9.755

Eddy SR (2002*a*) A memory efficient dynamic programming algorithm for optimal structural alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**, 18. doi: 10.1186/1471-2105-3-18

Eddy SR (2002*b*) Computational genomics of noncoding RNA genes. *Cell* **109**, 137–140. doi: 10.1016/S0092-8674(02)00727-4

Edgar RC (2004*a*) Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Research* **32**, 380–385. doi: 10.1093/nar/gkh180

Edgar RC (2004*b*) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797. doi: 10.1093/nar/gkh340

Edgar RC (2004*c*) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113. doi: 10.1186/1471-2105-5-113

Edgar RC, Sjölander K (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* **20**, 1301–1308. doi: 10.1093/bioinformatics/bth090

Edgar RC, Batzoglou S (2006) Multiple sequence alignment. *Current Opinion in Structural Biology* **16**, 368–373. doi: 10.1016/j.sbi.2006.04.004

Elias I (2003) Settling the intractability of multiple alignment. *Lecture Notes in Computer Science* **2906**, 352–363.

Ellis J, Morrison D (1995) Effects of sequence alignment on the phylogeny of *Sarcocystis* deduced from 18S rDNA sequences. *Parasitology Research* **81**, 696–699. doi: 10.1007/BF00931849

Errami M, Geourjon C, Deléage G (2003) Conservation of amino acids into multiple alignments involved in pairwise interactions in three-dimensional protein structures. *Journal of Bioinformatics and Computational Biology* **1**, 505–520. doi: 10.1142/S0219720003000228

Feng D-F, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**, 351–360.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Research* **34**, D247–D251. doi: 10.1093/nar/gkj149

Fitch WM (2000) Homology: a personal view on some of the problems. *Trends in Genetics* **16**, 227–231. doi: 10.1016/S0168-9525(00)02005-9

Fitch WM, Smith TF (1983) Optimal sequence alignments. *Proceedings of the National Academy of Sciences USA* **80**, 1382–1386. doi: 10.1073/pnas.80.5.1382

Fleißner R (2004) 'Sequence alignment and phylogenetic inference.' (Logos Verlag: Berlin)

Fleissner R, Metzler D, von Haeseler A (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology* **54**, 548–561. doi: 10.1080/10635150590950371

Frith MC, Hansen U, Spouge JL, Weng Z (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research* **32**, 189–200. doi: 10.1093/nar/gkh169

Gagnon S, Bourbeau D, Levesque RC (1996) Secondary structures and features of the 18S, 5.8S and 26S ribosomal RNAs from the Apicomplexan parasite *Toxoplasma gondii. Gene* **173**, 129–135. doi: 10.1016/0378-1119(96)00215-6

Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5**, 140. doi: 10.1186/1471-2105-5-140

Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research* **33**, 2433–2439. doi: 10.1093/nar/gki541

Geiger DL (2002) Stretch coding and block coding: two new strategies to represent questionably aligned DNA sequences. *Journal of Molecular Evolution* **54**, 191–199. doi: 10.1007/s00239-001-0001-5

Gille C, Frömmel C (2001) STRAP: editor for structural alignments of proteins. *Bioinformatics* **17**, 377–378. doi: 10.1093/bioinformatics/17.4.377

Gillespie JJ (2004) Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules. *Molecular Phylogenetics and Evolution* **33**, 936–943. doi: 10.1016/j.ympev.2004.08.004

Gillespie JJ, Yoder MJ, Wharton RA (2005*a*) Predicted secondary structure for 28S and 18S rRNA from Ichneumonoidea (Insecta: Hymenoptera: Apocrita): impact on sequence alignment and phylogeny estimation. *Journal of Molecular Evolution* **61**, 114–137. doi: 10.1007/s00239-004-0246-x

Gillespie JJ, McKenna CH, Yoder MJ, Gutell RR, Johnston JS, Kathirithamby J, Cognato AI (2005*b*) Assessing the odd secondary structural properties of nuclear small subunit ribosomal RNA sequences (18S) of the twisted-wing parasites (Insecta: Strepsiptera). *Insect Molecular Biology* **14**, 625–643. doi: 10.1111/j.1365-2583.2005.00591.x

Giribet G (2001) Exploring the behavior of POY, a program for direct optimization of molecular data. *Cladistics* **17**, S60–S70. doi: 10.1111/j.1096-0031.2001.tb00105.x

Giribet G (2002) Relationship among metazoan phyla as inferred from 18S rRNA sequence data: a methodological approach. In 'Molecular systematics and evolution: theory and practice'. (Eds R DeSalle, G Giribet, W Wheeler) pp. 85–101. (Birkhäuser Verlag: Basel)

Giribet G (2005) Generating implied alignments under direct optimization using POY. *Cladistics* **21**, 396–402. doi: 10.1111/j.1096-0031.2005.00071.x

Giribet G, Wheeler WC (1999) On gaps. *Molecular Phylogenetics and Evolution* **13**, 132–143. doi: 10.1006/mpev.1999.0643

Giribet G, Wheeler WC, Muona J (2002) DNA multiple sequence alignments. In 'Molecular systematics and evolution: theory and practice'. (Eds R DeSalle, G Giribet, W Wheeler) pp. 107–114. (Birkhäuser Verlag: Basel)

Gonnet GH, Korostensky C, Benner S (2000) Evaluation measures of multiple sequence alignments. *Journal of Computational Biology* **7**, 261–276. doi: 10.1089/10665270050081513

Gotoh O (1982) An improved algorithm for matching biological sequences. *Journal of Molecular Biology* **162**, 705–708. doi: 10.1016/0022-2836(82)90398-9

Gotoh O (1990) Consistency of optimal sequence alignments. *Bulletin of Mathematical Biology* **52**, 509–525.

Gotoh O (1995) A weighting scheme and algorithm for aligning many phylogenetically related sequences. *Computer Applications in the Biosciences* **11**, 543–551.

Gotoh O (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *Journal of Molecular Biology* **264**, 823–838. doi: 10.1006/jmbi.1996.0679

Gotoh O (1999) Multiple sequence alignment: algorithms and applications. *Advances in Biophysics* **36**, 159–206. doi: 10.1016/S0065-227X(99)80007-0

Gough J (2005) Convergent evolution of domain architectures is rare. *Bioinformatics* **21**, 1464–1471. doi: 10.1093/bioinformatics/bti204

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. *International Journal of Plant Sciences* **161**, S83–S96. doi: 10.1086/317583

Grasso C, Lee C (2004) Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics* **20**, 1546–1556. doi: 10.1093/bioinformatics/bth126

Greenberg HJ, Hart WE, Lancia G (2004) Opportunities for combinatorial optimization in computational biology. *INFORMS Journal on Computing* **16**, 211–231. doi: 10.1287/ijoc.1040.0073

Griffiths-Jones S (2005) RALEE—RNA alignment editor in emacs. *Bioinformatics* **21**, 257–259. doi: 10.1093/bioinformatics/bth489

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121–D124. doi: 10.1093/nar/gki081

Gu X, Li W-H (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *Journal of Molecular Evolution* **40**, 464–473. doi: 10.1007/BF00164032

Gueneau de Novoa P, Williams KP (2004) The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. *Nucleic Acids Research* **32**, D104–D108. doi: 10.1093/nar/gkh102

Gupta SK, Kececioglu JD, Schäffer AA (1995) Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *Journal of Computational Biology* **2**, 459–472.

Gutell RR, Lee JC, Cannone JJ (2002) The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology* **12**, 301–310. doi: 10.1016/S0959-440X(02)00339-1

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95–98.

Hancock JM, Vogler AP (2000) How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Molecular Phylogenetics and Evolution* **14**, 366–374. doi: 10.1006/mpev.1999.0709

Haszprunar G (1998) Parsimony analysis as a specific kind of homology estimation and the implications for character weighting. *Molecular Phylogenetics and Evolution* **9**, 333–339. doi: 10.1006/mpev.1998.0496

Heger A, Holm L (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function, and Genetics* **41**, 224–237. doi: 10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z

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

Hein J (1994) An algorithm combining DNA and protein alignment. *Journal of Theoretical Biology* **167**, 169–174. doi: 10.1006/jtbi.1994.1062

Hein J, Støvlbæk J (1996) Combined DNA and protein alignment. *Methods in Enzymology* **266**, 402–418.

Helm M, Brulé H, Friede D, Giegé R, Pütz J, Florentz C (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA* **6**, 1356–1379. doi: 10.1017/S1355838200001047

Henneke CM (1989) A multiple sequence alignment algorithm for homologous proteins using secondary structure information and optionally keying alignments to functionally important sites. *Computer Applications in the Biosciences* **5**, 141–150.

Hennig W (1966) 'Phylogenetic systematics.' [Transl. DD Davis, R Zangerl from W Hennig (1950) 'Grundzüge einer theorie der phylogenetischen systematik.' (Deutscher Zentralverlag: Berlin)] (University of Illinois Press: Urbana)

Henikoff S (1991) Playing with blocks: some pitfalls of forcing multiple alignments. *The New Biologist* **3**, 1148–1154.

Heringa J (1999) Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Computers and Chemistry* **23**, 341–364. doi: 10.1016/S0097-8485(99)00012-1

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. *Molecular Biology and Evolution* **13**, 150–169.

Hickson RE, Simon C, Perrey SW (2000) The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Molecular Biology and Evolution* **17**, 530–539.

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* **266**, 383–402.

Higgins DG, Blackshields G, Wallace IM (2005) Mind the gaps: progress in progressive alignment. *Proceedings of the National Academy of Sciences USA* **102**, 10 411–10 412. doi: 10.1073/pnas.0504801102

Higgs PG (2000) RNA secondary structure: physical and computational aspects. *Quarterly Reviews of Biophysics* **33**, 199–253. doi: 10.1017/S0033583500003620

Hillis DM (1994) Homology in molecular biology. In 'Homology: the hierarchical basis of comparative biology'. (Ed. BK Hall) pp. 339–368. (Academic Press: San Diego)

Hirosawa M, Totoki Y, Hoshida M, Ishikawa M (1995) Comprehensive study of iterative algorithms of multiple sequence alignment. *Computer Applications in the Biosciences* **11**, 13–18.

Hofacker IL, Bernhart SHF, Stadler PF (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics* **20**, 2222–2227. doi: 10.1093/bioinformatics/bth229

Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of Molecular Evolution* **20**, 175–186. doi: 10.1007/BF02257378

Holm L, Sander C (1996) Mapping the protein universe. *Science* **273**, 595–603. doi: 10.1126/science.273.5275.595

Holmes I (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* **6**, 73. doi: 10.1186/1471-2105-6-73

Holmes I, Durbin R (1998) Dynamic programming alignment accuracy. *Journal of Computational Biology* **5**, 493–504.

Hoot SB, Douglas AW (1998) Phylogeny of the Proteaceae based on *atpB* and *atpB–rbcL* intergenic spacer region sequences. *Australian Systematic Botany* **11**, 301–320. doi: 10.1071/SB98027

Hua Y, Jiang T, Wu B (1999) Aligning DNA sequences to minimize the change in protein. *Journal of Combinatorial Optimization* **3**, 227–245. doi: 10.1023/A:1009889710983

Huang X, Miller W (1991) A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* **12**, 337–357. doi: 10.1016/0196-8858(91)90017-D

Hudak J, McClure MA (1999) A comparative analysis of computational motif-detection methods. In 'Proceedings of the 4th Pacific Symposium on Biocomputing 1999, Hawaii'. pp. 138–149.

Janies DA, Wheeler WC (2001) Efficiency of parallel direct optimization. *Cladistics* **17**, S71–S82. doi: 10.1111/j.1096-0031.2001.tb00106.x

Jennings AJ, Edge CM, Sternberg MJE (2001) An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Engineering* **14**, 227–231. doi: 10.1093/protein/14.4.227

Jeon Y-S, Chung H, Park S, Hur I, Lee J-H, Chun J (2005) jPHYDIT: a JAVA-based integrated environment for molecular phylogeny of ribosomal RNA sequences. *Bioinformatics* **21**, 3171–3173. doi: 10.1093/bioinformatics/bti463

Jiang T, Lawler EL, Wang L (1994) Aligning sequences via an evolutionary tree: complexity and approximation. In 'Proceedings of the 26th annual ACM symposium on theory of computing'. pp. 760–769. (ACM Press: New York)

Johnson MS, Sali A, Blundell TL (1990) Phylogenetic relationships from three-dimensional protein structures. *Methods in Enzymology* **183**, 670–690.

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

Just W (2001) Computational complexity of multiple sequence alignment with SP-score. *Journal of Computational Biology* **8**, 615–623. doi: 10.1089/106652701753307511

Just W, Della Vedova G (2004) Multiple sequence alignment as a facility location problem. *INFORMS Journal on Computing* **16**, 430–440. doi: 10.1287/ijoc.1040.0093

Karaca M, Bilgen M, Onus AN, Ince AG, Elmasulu SY (2005) Exact Tandem Repeats Analyzer (E-TRA): a new program for DNA sequence mining. *Journal of Genetics* **84**, 49–54.

Karp RM (2002) Mathematical challenges from genomics and molecular biology. *Notices of the AMS* **49**, 544–553.

Karplus K, Hu B (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics* **17**, 713–720. doi: 10.1093/bioinformatics/17.8.713

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research* **30**, 3059–3066. doi: 10.1093/nar/gkf436

Katoh K, Kuma K, Toh H, Miyata T (2005*a*) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511–518. doi: 10.1093/nar/gki198

Katoh K, Kuma K, Miyata T, Toh H (2005*b*) Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Informatics* **16**, 22–33.

Kawakita A, Sota T, Ascher JS, Ito M, Tanaka H, Kato M (2003) Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble genes (*Bombus*). *Molecular Biology and Evolution* **20**, 87–92. doi: 10.1093/molbev/msg007

Kececioglu J, Starrett D (2004) Aligning alignments exactly. In 'Proceedings of the 8th ACM conference on research in computational molecular biology (RECOMB'04)'. pp. 85–96. (ACM Press: New York)

Kececioglu J, Kim E (2006) Simple and fast inverse alignment. *Lecture Notes in Computer Science* **3909**, 441–455.

Keightley PD, Johnson T (2004) MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. *Genome Research* **14**, 442–450. doi: 10.1101/gr.1571904

Kelchner SA (2000) The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* **87**, 482–498. doi: 10.2307/2666142

Kelchner SA (2002) Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *American Journal of Botany* **89**, 1651–1669.

Kelchner SA, Wendel JF (1996) Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Current Genetics* **30**, 259–262. doi: 10.1007/s002940050130

Kelchner SA, Clark LG (1997) Molecular evolution and phylogenetic utility of the chloroplast *rpl16* intron in *Chusquea* and the Bambusoideae (Poaceae). *Molecular Phylogenetics and Evolution* **8**, 385–397. doi: 10.1006/mpev.1997.0432

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. *Molecular Phylogenetics and Evolution* **4**, 314–330. doi: 10.1006/mpev.1995.1028

Kjer KM (1997) An alignment template for amphibian 12S rRNA, domain III: conserved primary and secondary structural motifs. *Journal of Herpetology* **31**, 599–604. doi: 10.2307/1565621

Kjer KM (2004) Aligned 18S and insect phylogeny. *Systematic Biology* **53**, 506–514. doi: 10.1080/10635150490445922

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

Kjer KM, Gillespie JJ, Ober KA (2006) Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. *Systematic Biology* (in press).

Kleinjung J, Douglas N, Heringa J (2002) Parallelized multiple alignment. *Bioinformatics* **18**, 1270–1271. doi: 10.1093/bioinformatics/18.9.1270

Knudsen B, Miyamoto M (2003) Sequence alignments and pair hidden markov models using evolutionary history. *Journal of Molecular Biology* **333**, 453–460. doi: 10.1016/j.jmb.2003.08.015

Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of Molecular Biology* **346**, 1173–1188. doi: 10.1016/j.jmb.2004.12.032

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature* **304**, 412–417. doi: 10.1038/304412a0

Kroken S, Taylor JW (2001) Outcrossing and recombination in the lichenized fungus *Letharia. Fungal Genetics and Biology* **34**, 83–92. doi: 10.1006/fgbi.2001.1291

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**, 426–427. doi: 10.1093/bioinformatics/15.5.426

Lambert C, Van Campenhout J-M, DeBolle X, Depiereux E (2003) Review of common sequence alignment methods: clues to enhance reliability. *Current Genomics* **4**, 131–146. doi: 10.2174/1389202033350038

Lancia G, Ravi R (1999) GESTALT: genomic steiner aligments. *Lecture Notes in Computer Science* **1645**, 101–114.

Lassmann T, Sonnhammer ELL (2002) Quality assessment of multiple alignment programs. *FEBS Letters* **529**, 126–130. doi: 10.1016/S0014-5793(02)03189-7

Lassmann T, Sonnhammer ELL (2005) Automatic assessment of alignment quality. *Nucleic Acids Research* **33**, 7120–7128. doi: 10.1093/nar/gki1020

Laurenne NM, Broad GR, Quicke DLJ (2006) Direct optimization and multiple alignment of 28S D2–D3 rDNA sequences: problems with indels on the way to a molecular phylogeny of the cryptine ichneumon wasps (Insecta: Hymenoptera). *Cladistics* **22**, 442–473. doi: 10.1111/j.1096-0031.2006.00112.x

Lawrence CJ, Malmberg RL, Muszynski MG, Dawe RK (2002) Maximum likelihood methods reveal conservation of function among closely related kinesin families. *Journal of Molecular Evolution* **54**, 42–53. doi: 10.1007/s00239-001-0016-y

Lawrence CJ, Zmasek CM, Dawe RK, Malmberg RL (2004) LumberJack: a heuristic tool for sequence alignment exploration and phylogenetic inference. *Bioinformatics* **20**, 1977–1979. doi: 10.1093/bioinformatics/bth180

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. *Molecular Biology and Evolution* **23**, 1180–1191. doi: 10.1093/molbev/msk010

Lecompte O, Thompson JD, Plewniak F, Thierry J-C, Poch O (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* **270**, 17–30. doi: 10.1016/S0378-1119(01)00461-9

Lee MSY (2001) Unalignable sequences and molecular evolution. *Trends in Ecology and Evolution* **16**, 681–685. doi: 10.1016/S0169-5347(01)02313-8

Lenhof H-P, Reinert K, Vingron M (1998) A polyhedral approach to RNA sequence structure alignment. *Journal of Computational Biology* **5**, 517–530.

Li K-B (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19**, 1585–1586. doi: 10.1093/bioinformatics/btg192

Lombard V, Camon EB, Parkinson HE, Hingamp P, Stoesser G, Redaschi N (2002) EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. *Bioinformatics* **18**, 763–764. doi: 10.1093/bioinformatics/18.5.763

Löytynoja A, Milinkovitch MC (2001) SOAP, cleaning multiple alignments from unstable blocks. *Bioinformatics* **17**, 573–574. doi: 10.1093/bioinformatics/17.6.573

Löytynoja A, Milinkovitch MC (2003) A hidden markov model for progressive multiple alignment. *Bioinformatics* **19**, 1505–1513. doi: 10.1093/bioinformatics/btg193

Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences USA* **102**, 10 557–10 562. doi: 10.1073/pnas.0409137102

Lu CL, Huang YP (2005) A memory-efficient algorithm for multiple sequence alignment with constraints. *Bioinformatics* **21**, 23–30.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, *et al*. (2004) ARB: a software environment for sequence data. *Nucleic Acids Research* **32**, 1363–1371. doi: 10.1093/nar/gkh293

Lunter G, Drummond AJ, Miklós I, Hein J (2005) Statistical alignment: recent progress, new applications, and challenges. In 'Statistical methods in molecular evolution'. (Ed. R Nielsen) pp. 375–405. (Springer: New York)

Manohar A, Batzoglou S (2005) TreeRefiner: a tool for refining a multiple alignment on a phylogenetic tree. In 'Proceedings of the 2005 IEEE computational systems bioinformatics conference (CSB'05)'. pp. 111–119. (IEEE Press: Piscataway)

Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH (2002) Comparison of sequence and structure alignments for protein domains. *Proteins: Structure, Function, and Genetics* **48**, 439–446. doi: 10.1002/prot.10163

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, *et al*. (2005) CDD: a conserved domain database for protein classification. *Nucleic Acids Research* **33**, D192–D196. doi: 10.1093/nar/gki069

Margulies EH, Chen CW, Green ED (2006) Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends in Genetics* **22**, 187–193. doi: 10.1016/j.tig.2006.02.005

Marsden B, Abagyan R (2004) SAD—a normalized structural alignment database: improving sequence–structure alignments. *Bioinformatics* **20**, 2333–2344. doi: 10.1093/bioinformatics/bth244

Marti-Renom MA, Madhusudhan MS, Sali A (2004) Alignment of protein sequences by their profiles. *Protein Science* **13**, 1071–1087. doi: 10.1110/ps.03379804

May ACW (2004) Percent sequence identity: the need to be explicit. *Structure* **12**, 737–738. doi: 10.1016/j.str.2004.04.001

McClure MA, Vasi TK, Fitch WM (1994) Comparative analysis of multiple protein-sequence alignment methods. *Molecular Biology and Evolution* **11**, 571–592.

Mecham J, Clement M, Snell Q, Freestone T, Seppi K, Crandall K (2006) Jumpstarting phylogenetic analysis. *International Journal of Bioinformatics Research and Applications* **2**, 19–35.

Miklós I, Lunter GA, Holmes I (2004) A "long indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution* **21**, 529–540. doi: 10.1093/molbev/msh043

Milinkovitch MC, LeDuc RG, Adachi J, Farnir F, Georges M, Hasegawa M (1996) Effects of character weighting and species sampling on phylogeny reconstruction: a case study based on DNA sequence data in cetaceans. *Genetics* **144**, 1817–1833.

Miller W (2001) Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17**, 391–397. doi: 10.1093/bioinformatics/17.5.391

Mindell DP (1991) Aligning DNA sequences: homology and phylogenetic weighting. In 'Phylogenetic analysis of DNA sequences'. (Eds MM Miyamoto, J Cracraft) pp. 73–89. (Oxford University Press: New York)

Morell V (1996) TreeBASE: the roots of phylogeny. *Science* **273**, 569. doi: 10.1126/science.273.5275.569

Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218. doi: 10.1093/bioinformatics/15.3.211

Morgenstern B, Prohaska SJ, Pohler D, Stadler PF (2006) Multiple sequence alignment with user-defined anchor points. *Algorithms for Molecular Biology* **1**, 6. doi: 10.1186/1748-7188-1-6

Morris P, Cobabe E (1991) Cuvier meets Watson and Crick: the utility of molecules as classical homologies. *Biological Journal of the Linnean Society* **44**, 307–324.

Morrison DA (2006) Phylogenetic analyses of parasites in the new millennium. *Advances in Parasitology* **63**, 1–124.

Morrison DA, Ellis JT (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. *Molecular Biology and Evolution* **14**, 428–441.

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. *Molecular Biology and Evolution* **17**, 1842–1853.

Myers G, Selznick S, Zhang Z, Miller W (1996) Progressive multiple alignment with constraints. *Journal of Computational Biology* **3**, 563–572.

Nguyen HD, Yoshihara I, Yamamori K, Yasunaga M (2002) Aligning multiple protein sequences by parallel hybrid genetic algorithm. *Genome Informatics* **13**, 123–132.

Nicholas HB, Ropelewski AJ, Deerfield DW (2002) Strategies for multiple sequence alignment. *BioTechniques* **32**, 572–591.

Notredame C (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* **3**, 131–144. doi: 10.1517/14622416.3.1.131

Notredame C, O'Brien EA, Higgins DG (1997) RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Research* **25**, 4570–4580. doi: 10.1093/nar/25.22.4570

Notredame C, Holm L, Higgins DG (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* **14**, 407–422. doi: 10.1093/bioinformatics/14.5.407

Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205–217. doi: 10.1006/jmbi.2000.4042

Nozaki Y, Bellgard M (2005) Statistical evaluation and comparison of a pairwise alignment algorithm that a priori assigns the number of gaps rather than employing gap penalties. *Bioinformatics* **21**, 1421–1428. doi: 10.1093/bioinformatics/bti198

O'Brien EA, Notredame C, Higgins DG (1998) Optimization of ribosomal RNA profile alignments. *Bioinformatics* **14**, 332–341. doi: 10.1093/bioinformatics/14.4.332

O'Donnell K, Kistler HC, Tacke BK, Casper HH (2000) Gene genealogies reveal global phylogeographic structure and reproductive isolation among lineages of *Fusarium graminearum*, the fungus causing wheat scab. *Proceedings of the National Academy of Sciences USA* **97**, 7905–7910. doi: 10.1073/pnas.130193297

Ogden TH, Rosenberg MS (2006) Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology* **55**, 314–328. doi: 10.1080/10635150500541730

Ohlson T, Wallner B, Elofsson A (2004) Profile–profile methods provide improved fold recognition: a study of different profile–profile alignment methods. *Proteins: Structure, Function, and Bioinformatics* **57**, 188–197. doi: 10.1002/prot.20184

Oliver T, Schmidt B, Nathan D, Clemens R, Maskell D (2005) Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics* **21**, 3431–3432. doi: 10.1093/bioinformatics/bti508

Ophir R, Graur D (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**, 191–202. doi: 10.1016/S0378-1119(97)00398-3

O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology* **340**, 385–395. doi: 10.1016/j.jmb.2004.04.058

Page RDM (2000) Comparative analysis of secondary structure of insect mitochondrial small subunit ribosomal RNA using maximum weighted matching. *Nucleic Acids Research* **28**, 3839–3845. doi: 10.1093/nar/28.20.3839

Parida L, Floratos A, Rigoutsos I (1999) An approximation algorithm for alignment of multiple sequences using motif discovery. *Journal of Combinatorial Optimization* **3**, 247–275. doi: 10.1023/A:1009841927822

Parmentier G, Trystram D, Zola J (2004) Cache-based parallelization of multiple sequence alignment problem. *Lecture Notes in Computer Science* **3149**, 1005–1012.

Pascarella S, Argos P (1992) Analysis of insertions / deletions in protein structures. *Journal of Molecular Biology* **224**, 461–471. doi: 10.1016/0022-2836(92)91008-D

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

Pearson WR, Sierk ML (2005) The limits of protein sequence comparison? *Current Opinion in Structural Biology* **15**, 254–260. doi: 10.1016/j.sbi.2005.05.005

Pedersen CNS, Lyngsø R, Hein J (1998) Comparison of coding DNA. *Lecture Notes in Computer Science* **1448**, 153–173.

Pei J, Grishin NV (2006) MUMMALS: multiple sequence alignment improved by using hidden markov models with local structural information. *Nucleic Acids Research* **34**, 4364–4374. doi: 10.1093/nar/gkl514

Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics* **19**, 427–428. doi: 10.1093/bioinformatics/btg008

Petersen G, Seberg O, Aagesen L, Frederiksen S (2004) An empirical test of the treatment of indels during optimization alignment based on the phylogeny of the genus *Secale* (Poaceae). *Molecular Phylogenetics and Evolution* **30**, 733–742. doi: 10.1016/S1055-7903(03)00206-9

Pettersson EU, Ljunggren EL, Morrison DA, Mattsson JG (2005) Functional analysis and localisation of a class delta glutathione *S*-transferase from *Sarcoptes scabiei. International Journal for Parasitology* **35**, 39–48. doi: 10.1016/j.ijpara.2004.09.006

Phillips A (2006) Homology assessment and molecular sequence alignment. *Journal of Biomedical Informatics* **39**, 18–33. doi: 10.1016/j.jbi.2005.11.005

Phillips A, Janies D, Wheeler W (2000) Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution* **16**, 317–330. doi: 10.1006/mpev.2000.0785

Pible O, Imbert G, Pellequer J-L (2005) INTERALIGN: interactive alignment editor for distantly related protein sequences. *Bioinformatics* **21**, 3166–3167. doi: 10.1093/bioinformatics/bti474

de Pinna MCC (1991) Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7**, 367–394. doi: 10.1111/j.1096-0031.1991.tb00045.x

Poch O, Delarue M (1996) Converting sequence block alignments into structural insights. *Methods in Enzymology* **266**, 662–680.

Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**, 6. doi: 10.1186/1471-2105-5-6

Ponting CP, Birney E (2000) Identification of domains from protein sequences. In 'Protein structure prediction: methods and protocols'. (Ed. DM Webster) pp. 53–69. (Humana Press: Totowa)

Qian B, Goldstein RA (2001) Distribution of indel lengths. *Proteins: Structure, Function, and Genetics* **45**, 102–104. doi: 10.1002/prot.1129

Raghava GPS, Searle SMJ, Audley PC, Barber JD, Barton GJ (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* **4**, 47. doi: 10.1186/1471-2105-4-47

Rainaldi G, Volpicella M, Licciulli F, Liuni S, Gallerani R, Ceci LR (2003) PLMItRNA, a database on the heterogeneous genetic origin of mitochondrial tRNA genes and tRNAs in photosynthetic eukaryotes. *Nucleic Acids Research* **31**, 436–438. doi: 10.1093/nar/gkg080

Raphael B, Zhi D, Tang H, Pevzner P (2004) A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Research* **14**, 2336–2346. doi: 10.1101/gr.2657504

Redelings BD, Suchard MA (2005) Joint bayesian estimation of alignment and phylogeny. *Systematic Biology* **54**, 401–418. doi: 10.1080/10635150590947041

Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, *et al*. (1987) "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* **50**, 667. doi: 10.1016/0092-8674(87)90322-9

Reese JT, Pearson WR (2002) Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* **18**, 1500–1507. doi: 10.1093/bioinformatics/18.11.1500

Reinert K, Stoye J, Will T (2000) An iterative method for faster sum-of-pairs multiple sequence alignment. *Bioinformatics* **16**, 808–814. doi: 10.1093/bioinformatics/16.9.808

Riaz T, Wang Y, Li K-B (2004) Multiple sequence alignment using tabu search. *Conferences in Research and Practice in Information Technology* **29**, 223–232.

Riaz T, Wang Y, Li K-B (2005) Tabu search algorithm for post-processing multiple sequence alignment. *Journal of Bioinformatics and Computational Biology* **3**, 145–156. doi: 10.1142/S0219720005000928

Rice KA, Donoghue MJ, Olmstead RG (1997) Analyzing large data sets: rbcL 500 revisited. *Systematic Biology* **46**, 554–563. doi: 10.2307/2413696

Rieppel O (1994) Homology, topology, and typology: the history of modern debates. In 'Homology: the hierarchical basis of comparative biology'. (Ed. BK Hall) pp. 63–100. (Academic Press: San Diego)

Rieppel O, Kearney M (2002) Similarity. *Biological Journal of the Linnean Society* **75**, 59–82. doi: 10.1046/j.1095-8312.2002.00006.x

Rinsma-Melchert I (1993) The expected number of matches in optimal global sequence alignments. *New Zealand Journal of Botany* **31**, 219–230.

Rodriguez R, Vriend G (1997) Professional gambling. In 'Biomolecular structure and dynamics: recent experimental and theoretical advances'. (Eds G Vergoten, T Theophanides) pp. 79–120. (Kluwer Academic Publishers: Dordrecht)

Rosenberg MS (2005*a*) Evolutionary distance estimation and fidelity of pair wise sequence alignment. *BMC Bioinformatics* **6**, 102. doi: 10.1186/1471-2105-6-102

Rosenberg MS (2005*b*) MySSP: non-stationary evolutionary sequence simulation, including indels. *Evolutionary Bioinformatics Online* **1**, 81–83.

Roshan U, Livesay DR (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22**, 2715–2721. doi: 10.1093/bioinformatics/bt1472

Rost B, Valencia A (1996) Pitfalls of protein sequence analysis. *Current Opinion in Biotechnology* **7**, 457–461. doi: 10.1016/S0958-1669(96)80124-8

Sadreyev RI, Grishin NV (2004) Estimates of statistical significance for comparison of individual positions in multiple sequence alignments. *BMC Bioinformatics* **5**, 106. doi: 10.1186/1471-2105-5-106

Sammeth M, Heringa J (2006) Global multiple-sequence alignment with repeats. *Proteins: Structure, Function, and Bioinformatics* **64**, 263–274. doi: 10.1002/prot.20957

Sammeth M, Weniger T, Harmsen D, Stoye J (2005) Alignment of tandem repeats with excision, duplication, substitution and indels (EDSI). *Lecture Notes in Computer Science* **3692**, 276–290.

Sanchis A, Michelana JM, Latorre A, Quicke DLJ, Gärdenfors U, Belshaw R (2001) The phylogenetic analysis of variable-length sequence data: elongation factor-1α introns in European populations of the parasitoid wasp genus *Pauesia* (Hymenoptera: Braconidae: Aphidiinae). *Molecular Biology and Evolution* **18**, 1117–1131.

Sankoff D, Cedergren RJ (1983) Simultaneous comparison of three or more sequences related by a tree. In 'Time warps, string edits, and macromolecules: the theory and practice of sequence comparison'. (Eds D Sankoff, JB Kruskal) pp. 253–264. (Addison-Wesley: Reading)

Sankoff D, Morel C, Cedergren RJ (1973) Evolution of 5S RNA and the non-randomness of base replacement. *Nature* **245**, 232–234. doi: 10.1038/245232a0

Sauder JM, Arthur JW, Dunbrack RL (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Structure, Function, and Genetics* **40**, 6–22. doi: 10.1002/(SICI)1097-0134(20000701)40:1<6::AID-PROT30>3.0.CO;2-7

Schmollinger M, Nieselt K, Kaufmann M, Morgenstern B (2004) DIALIGN P: fast pair-wise and multiple sequence alignment using parallel processors. *BMC Bioinformatics* **5**, 128. doi: 10.1186/1471-2105-5-128

Schuler GD, Altschul SF, Lipman DJ (1991) A workbench for multiple alignment construction and analysis. *Proteins* **9**, 180–190. doi: 10.1002/prot.340090304

Schultes EA, Hraber PT, LaBean TH (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *Journal of Molecular Evolution* **49**, 76–83. doi: 10.1007/PL00006536

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. doi: 10.1261/rna.7204505

Schwikowski B, Vingron M (1997*a*) The deferred path heuristic for the generalized tree alignment problem. *Journal of Computational Biology* **4**, 415–431.

Schwikowski B, Vingron M (1997*b*) A clustering approach to generalized tree alignment with application to Alu repeats. *Lecture Notes in Computer Science* **1278**, 115–124.

Schwikowski B, Vingron M (2003) Sequence graphs: boosting iterated dynamic programming using locally suboptimal solutions. *Discrete Applied Mathematics* **127**, 95–117. doi: 10.1016/S0166-218X(02)00288-3

Shakhnovich BE (2005) Improving the precision of the structure–function relationship by considering phylogenetic context. *PLoS Computational Biology* **1**, e9. doi: 10.1371/journal.pcbi.0010009

Shull VL, Vogler AP, Baker MD, Maddison DR, Hammond PM (2001) Sequence alignment of 18S ribosomal RNA and the basal relationships of adephagan beetles: evidence for monophyly of aquatic families and the placement of Trachypachidae. *Systematic Biology* **50**, 945–969. doi: 10.1080/106351501753462894

Siddharthan R (2006) Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. *BMC Bioinformatics* **7**, 143. doi: 10.1186/1471-2105-7-143

Siebert S, Backofen R (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* **21**, 3352–3359. doi: 10.1093/bioinformatics/bti550

Simmons MP (2004) Independence of alignment and tree search. *Molecular Phylogenetics and Evolution* **31**, 874–879. doi: 10.1016/j.ympev.2003.10.008

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analysis. *Systematic Biology* **49**, 369–381. doi: 10.1080/10635159950173889

Simmons MP, Freudenstein JV (2003) The effects of increasing genetic distance on alignment of, and tree construction from, rDNA internal transcribed spacer sequences. *Molecular Phylogenetics and Evolution* **26**, 444–451. doi: 10.1016/S1055-7903(02)00366-4

Simmons MP, Carr TG, O'Neill K (2004) Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. *Molecular Phylogenetics and Evolution* **32**, 913–926. doi: 10.1016/j.ympev.2004.04.011

Simossis VA, Heringa J (2004) Integrating protein secondary structure prediction and multiple sequence alignment. *Current Protein and Peptide Science* **5**, 249–266. doi: 10.2174/1389203043379675

Simossis VA, Heringa J (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research* **33**, W289–W294. doi: 10.1093/nar/gki390

Simossis VA, Kleinjung J, Heringa J (2005) Homology-extended sequence alignment. *Nucleic Acids Research* **33**, 816–824. doi: 10.1093/nar/gki233

Slowinski JB (1998) The number of multiple alignments. *Molecular Phylogenetics and Evolution* **10**, 264–266. doi: 10.1006/mpev.1998.0522

Sluys R (1996) The notion of homology in current comparative biology. *Journal of Zoological Systematics and Evolutionary Research* **34**, 145–152.

Smith NGC, Hurst LD (1998) Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. *Journal of Molecular Evolution* **47**, 493–500. doi: 10.1007/PL00013151

del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. *Journal of Molecular Biology* **326**, 1289–1302. doi: 10.1016/S0022-2836(02)01451-1

Sprinzl M, Vassilenko KS (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research* **33**, D139–D140. doi: 10.1093/nar/gki012

Stebbings LA, Mizuguchi K (2004) HOMSTRAD: recent developments of the homologous protein structure alignment database. *Nucleic Acids Research* **32**, D203–D207. doi: 10.1093/nar/gkh027

Stocsits RR, Hofaker IL, Fried C, Stadler PF (2005) Multiple sequence alignments of partially coding nucleic acid sequences. *BMC Bioinformatics* **6**, 160. doi: 10.1186/1471-2105-6-160

Stoye J, Evers D, Meyer F (1998) Rose: generating sequence families. *Bioinformatics* **14**, 157–163. doi: 10.1093/bioinformatics/14.2.157

Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics* **6**, 66. doi: 10.1186/1471-2105-6-66

Sze S-H, Lu Y, Yang Q (2006) A polynomial time solvable formulation of multiple sequence alignment. *Journal of Computational Biology* **13**, 309–319. doi: 10.1089/cmb.2006.13.309

Szklarczyk R, Heringa J (2004) Tracking repeats using significance and transitivity. *Bioinformatics* **20**, i311–i317. doi: 10.1093/bioinformatics/bth911

Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J (2002) 5S ribosomal RNA database. *Nucleic Acids Research* **30**, 176–178. doi: 10.1093/nar/30.1.176

Taylor WR (1986) Identification of protein sequence homology by consensus template alignment. *Journal of Molecular Biology* **188**, 233–258. doi: 10.1016/0022-2836(86)90308-6

Taylor WR (1987) Multiple sequence alignment by a pairwise algorithm. *Computer Applications in the Biosciences* **3**, 81–87.

Taylor WR (1996) Multiple protein sequence alignment: algorithms and gap insertion. *Methods in Enzymology* **266**, 343–367.

Teeling H, Gloeckner FO (2006) RibAlign: a software tool and database for eubacterial phylogeny based on concatenated ribosomal protein subunits. *BMC Bioinformatics* **7**, 66. doi: 10.1186/1471-2105-7-66

Telford MJ, Wise MJ, Gowri-Shankar V (2005) Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the Bilateria. *Molecular Biology and Evolution* **22**, 1129–1136. doi: 10.1093/molbev/msi099

Terry MD, Whiting MF (2005) Comparison of two alignment techniques within a single complex data set: POY versus Clustal. *Cladistics* **21**, 272–281. doi: 10.1111/j.1096-0031.2005.00063.x

Thébault P, Monestié A, Higgins DG (1999) MIAH: automatic alignment of eukaryotic SSU rRNAs. *Bioinformatics* **15**, 341–342. doi: 10.1093/bioinformatics/15.4.341

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 Research* **22**, 4673–4680.

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 Research* **25**, 4876–4882. doi: 10.1093/nar/25.24.4876

Thompson JD, Plewniak F, Poch O (1999*a*) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**, 87–88. doi: 10.1093/bioinformatics/15.1.87

Thompson JD, Plewniak F, Poch O (1999*b*) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research* **27**, 2682–2690. doi: 10.1093/nar/27.13.2682

Thompson JD, Plewniak F, Thierry J-C, Poch O (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Research* **28**, 2919–2926. doi: 10.1093/nar/28.15.2919

Thompson JD, Plewniak F, Ripp R, Thierry J-C, Poch O (2001) Towards a reliable objective function for multiple sequence alignments. *Journal of Molecular Biology* **314**, 937–951. doi: 10.1006/jmbi.2001.5187

Thompson JD, Thierry JC, Poch O (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics* **19**, 1155–1161. doi: 10.1093/bioinformatics/btg133

Thompson JD, Koehl P, Ripp R, Poch O (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics* **61**, 127–136. doi: 10.1002/prot.20527

Thomsen R, Fogel GB, Krink T (2002) A Clustal alignment improver using evolutionary algorithms. In 'Proccedings of the fourth congress on evolutionary computation (CEC-2002)'. (Eds DB Fogel, X Yao, G Greenwood, H Iba, P Marrow, M Shackleton) pp. 121–126. (IEEE Press: Piscataway)

Thomsen R, Fogel GB, Krink T (2003) Improvement of Clustal-derived sequence alignments with evolutionary algorithms. In 'Proccedings of the fifth congress on evolutionary computation (CEC-2003)'. (Eds DR Sarker, R Reynolds, H Abbass, KC Tan, B McKay, D Essam, T Gedeon) pp. 1499–1507. (IEEE Press: Piscataway)

Thorne JL, Kishino H (1992) Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution* **9**, 1148–1162.

Thorne JL, Churchill GA (1995) Estimation and reliability of molecular sequence alignments. *Biometrics* **51**, 100–113. doi: 10.2307/2533318

Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**, 114–124. doi: 10.1007/BF02193625

Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality: an improved likelihood model for sequence evolution. *Journal of Molecular Evolution* **34**, 3–16. doi: 10.1007/BF00163848

Titus TA, Frost DR (1996) Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). *Molecular Phylogenetics and Evolution* **6**, 49–62. doi: 10.1006/mpev.1996.0057

Touzet H, Perriquet O (2004) CARNAC: folding families of related RNAs. *Nucleic Acids Research* **32**, W142–W145.

Trystram D, Zola J (2005) Parallel multiple sequence alignment with decentralized cache support. *Lecture Notes in Computer Science* **3648**, 1217–1226.

Tsai YT, Huang YP, Yu CT, Lu CL (2004) MuSiC: a tool for multiple sequence alignment with constraints. *Bioinformatics* **20**, 2309–2311. doi: 10.1093/bioinformatics/bth220

Tyson H (1992) Relationships between amino acid sequences determined through optimum alignments, clustering, and specific distance patterns: application to a group of scorpion toxins. *Genome* **35**, 360–371.

van Valen L (1982) Homology and causes. *Journal of Morphology* **173**, 305–312. doi: 10.1002/jmor.1051730307

Van Walle I, Lasters I, Wyns L (2004) Align-m—a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics* **20**, 1428–1435. doi: 10.1093/bioinformatics/bth116

Van Walle I, Lasters I, Wyns L (2005) SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **21**, 1267–1268. doi: 10.1093/bioinformatics/bth493

Varani G, Pardi A (1994) Structure of RNA. In 'RNA–protein interactions'. (Eds K Nagai, IW Mattaj) pp. 1–24. (IRL Press: Oxford)

Vingron M (1999) Sequence alignment and phylogeny construction. In 'Mathematical support for molecular biology'. (Eds M Farach-Colton, FS Roberts, M Vingron, M Waterman) pp. 53–64. (American Mathematical Society: Providence)

Vingron M, Waterman MS (1994) Sequence alignments and penalty choice: review of concepts, case studies and implications. *Journal of Molecular Biology* **235**, 1–12. doi: 10.1016/S0022-2836(05)80006-3

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

Vogt G, Etzold T, Argos P (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *Journal of Molecular Biology* **249**, 816–831. doi: 10.1006/jmbi.1995.0340

Vogt L (2002) Testing and weighting characters. *Organisms, Diversity and Evolution* **2**, 319–333. doi: 10.1078/1439-6092-00051

Wagner GP (1989) The biological homology concept. *Annual Review of Ecology and Systematics* **20**, 51–69. doi: 10.1146/annurev.es.20.110189.000411

Wallace IM, Blackshields G, Higgins DG (2005*a*) Multiple sequence alignments. *Current Opinion in Structural Biology* **15**, 261–266. doi: 10.1016/j.sbi.2005.04.002

Wallace IM, O'Sullivan O, Higgins DG (2005*b*) Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics* **21**, 1408–1414. doi: 10.1093/bioinformatics/bti159

Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research* **34**, 1692–1699. doi: 10.1093/nar/gkl091

Wang G, Dunbrack RL (2004) Scoring profile-to-profile sequence alignments. *Protein Science* **13**, 1612–1626. doi: 10.1110/ps.03601504

Wang L, Jiang T (1994) On the complexity of multiple sequence alignment. *Journal of Computational Biology* **1**, 337–348.

Wang Y, Li K-B (2004) An adaptive and iterative algorithm for refining multiple sequence alignment. *Computational Biology and Chemistry* **28**, 141–148. doi: 10.1016/j.compbiolchem.2004.02.001

Wareham HT (1995) A simplified proof of the NP- and MAX SNP-hardness of multiple sequence tree alignment. *Journal of Computational Biology* **2**, 509–514.

Waterman MS (1995) 'Introduction to computational biology: maps, sequences and genomes.' (Chapman & Hall: London)

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

Wegnez M (1987) Letter to the editor. *Cell* **51**, 516. doi: 10.1016/0092-8674(87)90118-8

Wernersson R, Pedersen AG (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research* **31**, 3537–3539. doi: 10.1093/nar/gkg609

Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Research* **31**, 489–491. doi: 10.1093/nar/gkg068

Wexler Y, Yakhini Z, Kashi Y, Geiger D (2005) Finding approximate tandem repeats in genomic sequences. *Journal of Computational Biology* **12**, 928–942. doi: 10.1089/cmb.2005.12.928

Wheeler WC (1993) The triangle inequality and character analysis. *Molecular Biology and Evolution* **10**, 707–712.

Wheeler WC (1994) Sources of ambiguity in nucleic acid sequence alignment. In 'Molecular ecology and evolution: approaches and applications'. (Eds B Schierwater, B Streit, GP Wagner, R DeSalle) pp. 323–352. (Birkhäuser Verlag: Basel)

Wheeler WC (1995) Sequence alignment, parameter sensitivity, and phylogenetic analysis of molecular data. *Systematic Biology* **44**, 321–331. doi: 10.2307/2413595

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

Wheeler W (1998) Alignment characters, dynamic programming and heuristic solutions. In 'Molecular approaches to ecology and evolution'. (Eds R DeSalle, B Schierwater) pp. 243–251. (Birkhäuser Verlag: Basel)

Wheeler WC (1999) Fixed character states and the optimization of molecular sequence data. *Cladistics* **15**, 379–385. doi: 10.1111/j.1096-0031.1999.tb00274.x

Wheeler W (2001*a*) Homology and DNA sequence data. In 'The character concept in evolutionary biology'. (Ed. GP Wagner) pp. 303–317. (Academic Press: San Diego)

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

Wheeler WC (2002) Optimization alignment: down, up, error, and improvements. In 'Techniques in molecular systematics and evolution'. (Eds R DeSalle, G Giribet, W Wheeler) pp. 55–69. (Birkhäuser Verlag: Basel)

Wheeler WC (2003*a*) Iterative pass optimization of sequence data. *Cladistics* **19**, 254–260. doi: 10.1111/j.1096-0031.2003.tb00368.x

Wheeler WC (2003*b*) Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. *Cladistics* **19**, 261–268. doi: 10.1111/j.1096-0031.2003.tb00369.x

Wheeler WC (2003*c*) Search-based optimization. *Cladistics* **19**, 348–355. doi: 10.1111/j.1096-0031.2003.tb00378.x

Wheeler WC (2005) Alignment, dynamic homology, and optimization. In 'Parsimony, phylogeny, and genomics'. (Ed. VA Albert) pp. 71–80. (Oxford University Press: Oxford)

Wheeler WC (2006) Dynamic homology and the likelihood criterion. *Cladistics* **22**, 157–170. doi: 10.1111/j.1096-0031.2006.00096.x

Wheeler WC, Gladstein DS (1994) MALIGN: a multiple sequence alignment program. *Journal of Heredity* **85**, 417–418.

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 Research* **34**, D327–D331. doi: 10.1093/nar/gkj087

Whiting AS, Sites JW, Pellegrino KCM, Rodrigues MT (2006) Comparing alignment methods for inferring the history of the new world lizard genus *Mabuya* (Squamata: Scincidae). *Molecular Phylogenetics and Evolution* **38**, 719–730. doi: 10.1016/j.ympev.2005.11.011

Williams DM (1993) A note on molecular homology: multiple patterns from single datasets. *Cladistics* **9**, 233–245. doi: 10.1111/j.1096-0031.1993.tb00221.x

Winnepenninckx B, Backeljau T (1996) 18S rRNA alignments derived from different secondary structure models can produce alternative phylogenies. *Journal of Zoological Systematics and Evolutionary Research* **34**, 135–143.

Winter WP, Walsh KA, Neurath H (1968) Homology as applied to proteins. *Science* **162**, 1433. doi: 10.1126/science.162.3861.1433

Wrabl JO, Grishin NV (2004) Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. *Proteins: Structure, Function, and Bioinformatics* **54**, 71–87. doi: 10.1002/prot.10508

Wuyts J, Perrière G, Van de Peer Y (2004) The European ribosomal RNA database. *Nucleic Acids Research* **32**, D101–D103. doi: 10.1093/nar/gkh065

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. *International Journal for Parasitology* **32**, 1773–1785. doi: 10.1016/S0020-7519(02)00197-2

Yamada S, Gotoh O, Yamana H (2004) Extension of Prrn: implementation of a doubly nested randomized iterative refinement strategy under a piecewise linear gap cost. *Genome Informatics* **15**, P082.

Yu H, Deng M (2005) ClustalY: speed up the guide tree building for ClustalW. In 'Proceedings of the eighth international conference on high-performance computing in Asia-Pacific region (HPCASIA'05)'. pp. 608–610. (IEEE Press: Piscataway)

Yuan J, Amend A, Borkowski J, DeMarco R, Bailey W, Liu Y, Xie G, Blevins R (1999) MULTICLUSTAL: a systematic method for surveying ClustalW alignment parameters. *Bioinformatics* **15**, 862–863. doi: 10.1093/bioinformatics/15.10.862

Zhang X, Kahveci T (2006) A new approach for alignment of multiple proteins. In 'Proceedings of the 11th Pacific Symposium on Biocomputing 2006, Hawaii'. pp. 339–350.

Zhou H, Zhou Y (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structure. *Bioinformatics* **21**, 3615–3621. doi: 10.1093/bioinformatics/bti582

Zhu J, Liu JS, Lawrence CE (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**, 25–39. doi: 10.1093/bioinformatics/14.1.25

Zwieb C (1997) The uRNA database. *Nucleic Acids Research* **25**, 102–103. doi: 10.1093/nar/25.1.102

Manuscript received 3 July 2006, accepted 30 October 2006