Multiple sequence alignment: In pursuit of homologous DNA positions

Sudhir Kumar and Alan Filipski

Genome Res. 2007 17: 127-135
Access the most recent version at doi:10.1101/gr.5232407

Review

Multiple sequence alignment: In pursuit of homologous DNA positions

Sudhir Kumar[1] and Alan Filipski

Center for Evolutionary Functional Genomics, Biodesign Institute and School of Life Sciences, Arizona State University, Tempe, Arizona 85287-5301, USA

DNA sequence alignment is a prerequisite to virtually all comparative genomic analyses, including the identification of conserved sequence motifs, estimation of evolutionary divergence between sequences, and inference of historical relationships among genes and species. While it is mere common sense that inaccuracies in multiple sequence alignments can have detrimental effects on downstream analyses, it is important to know the extent to which the inferences drawn from these alignments are robust to errors and biases inherent in all sequence alignments. A survey of investigations into strengths and weaknesses of sequence alignments reveals, as expected, that alignment quality is generally poor for two distantly related sequences and can often be improved by adding additional sequences as stepping stones between distantly related species. Errors in sequence alignment are also found to have a significant negative effect on subsequent inference of sequence divergence, phylogenetic trees, and conserved motifs. However, our understanding of alignment biases remains rudimentary, and sequence alignment procedures continue to be used somewhat like benign formatting operations to make sequences equal in length. Because of the central role these alignments now play in our endeavors to establish the tree of life and to identify important parts of genomes through evolutionary functional genomics, we see a need for increased community effort to investigate influences of alignment bias on the accuracy of large-scale comparative genomics.

The relative positions of nucleotides within the same gene in different species and in duplicated genomic regions are disturbed by insertion and deletion of stretches of DNA over evolutionary time. This leads to differences in the length of the homologous regions in the genome, with more distant relatives having a higher likelihood of sequence length difference. A comparison of lengths of genome segments spanning protein-coding genes in human and mouse shows the extent of the effect of evolution by insertions and deletions (Fig. 1). The lengths of noncoding orthologous sequences have also evolved substantially after divergence over 90 million years ago. A grand challenge in comparative genomics is to line up these bases by inserting gaps in sequences, because genomic analyses must be based on comparisons between bases at positions (sites) that coincided in a common ancestor. The task is to re-establish (estimate) the ancestral site-wise homology obfuscated by the insertion–deletion and substitution processes. Naturally, this operation has come to be known as "alignment," and the resulting set of sequences, all of which are the same length (taking gaps in to account), is also called an alignment (Fig. 2). We can distinguish between "pairwise" alignments, in which sequences, even if they are part of a larger set, are aligned only in pairs, and "multiple" alignments, in which more than two sequences are aligned simultaneously.

Alignment procedures may also be classified as either "global" or "local." In the simplest form, sequences are aligned beginning to end to produce global alignments. This is appropriate for sequences of protein-coding genes and for short stretches of the genomic sequence. For longer genomic DNA, it is necessary to account for medium and large-scale rearrangements in addition to large sequence insertions and deletions, which necessitates the building of local alignments. Local alignments differ from global alignments in that the former focus on shared regions of high similarity while ignoring regions that do not show high sequence homology between sequences. Unlike traditional global and local alignment methods that assume colinearity of homologous segments among sequences, "glocal" alignment methods model the rearrangement process explicitly during the alignment procedure itself and result in a nonlinear mapping between homologous regions of different sequences (Brudno et al. 2003b).

For most applications in the areas of molecular phylogenetics and evolution, we are interested in properties and relationships of "rows" of the alignment, which represent species, genes, or genomic regions (Fig. 2). Examples of these include inference of multigene family and species phylogenies, determination of evolutionary rates in different lineages, and identification of bouts of selection and patterns of DNA sequence change over time (Nei and Kumar 2000; Felsenstein 2002). The second use of sequence alignments (local as well as global) stems from the need to understand properties of individual or groups of contiguous "columns." A common example is the building of genomic profiles of conservation and divergence in different DNA positions or sets of contiguous positions (genomic regions). These applications are considered to fall in the area of functional genomics (Hardison 2000; Gaucher et al. 2002; Town 2002; Koonin and Galperin 2003; Pevsner 2003). Of course, many studies involve joint analyses of rows and columns, but from the perspective of accuracy, the requirements of these two alignment usages are quite different. In the following sections, we assess the existing knowledge of the fidelity of the alignment process, focusing primarily on finding motifs, estimating sequence divergence, and inferring phylogenetic relationships.

[1]Corresponding author.
E-mail s.kumar@asu.edu; fax (480) 727-6947.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.5232407.

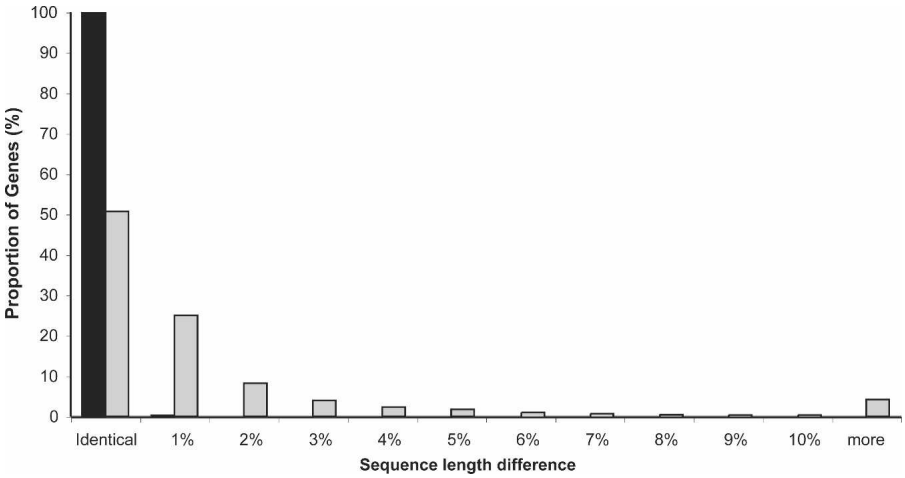

Figure 1. Frequency distributions of percent difference in total coding sequence length (exons only) between human and mouse orthologs (light gray) and between human and chimpanzee orthologs (black) for 7645 protein-coding genes. Forty-nine percent of the genes exhibit some length difference between human and mouse; human and chimpanzee orthologs exhibit length difference in only 17 of these 7645 genes. Data from Clark et al. (2003).

Practical sequence alignment

The amount of nucleotide sequence data in GenBank and other public databases has expanded exponentially since the inception of these electronic warehouses; the data now consist of over 125 billion base pairs of sequence data from over 200,000 organisms. Understanding the functional significance of these data has become the central problem in comparative genomics. Naturally, because of this great volume of data, scientists would like to establish DNA homologies by applying one or more of the highly innovative alignment methods available today in an automated high-throughput fashion (Thompson et al. 1994; Eddy 1995; Morgenstern et al. 1998; Notredame et al. 2000; Brudno et al. 2003a; Covert et al. 2004).

The process of alignment involves insertion of gaps into sequences to make them the same length. These gaps are hypotheses about the site homologies resulting from historical insertion–deletion events. Since mutations can cause two homologous sites to differ from each other (substitutions), the complexities of the alignment process transcend traditional string-matching problems in computer science. The interplay of insertion–deletion events and substitutions over thousands and millions of years produces sequences that may lead to many different alignments, with some containing more gaps than others. One may prefer a specific alignment over another if some "optimality" score is better (Needleman and Wunsch 1970). Such a score typically represents some function of the numbers and positions of columns that contain identical bases, different bases, or gaps. Each difference is penalized and there are penalties for inserting and extending gaps.

The scoring functions incorporate differences in the likelihood of change from one base to another and of inserting of gaps of various lengths. It is well-established, for example, that transitional mutations are much more common than transversional mutations (Vogel and Kopun 1977; Rosenberg and Kumar 2003). Hence, transitional differences are penalized less (i.e., they are allowed in the alignments more frequently). Longer gaps are also known to be rare as compared with shorter ones, so alignments containing long gaps are favored less (Waterman et al. 1976; Gu and Li 1995; Qian and Goldstein 2001; Miklos et al. 2004). However, the exact value to be assigned to each parameter is difficult to determine, as it is expected to differ between closely and distantly related species and vary from species to species (Vingron and Waterman 1994; Wheeler 1995; Zhu et al. 1998; Yuan et al. 1999). Although progress has been made in establishing biological justification for the use of some scoring schemes, a comprehensive theoretical framework for gap and substitution penalties still eludes us (Gu and Li 1995; Miklos et al. 2004). Thus, the optimal alignment under a specific scoring scheme may not be the true one (Landan 2005; Morgenstern et al. 2006). In fact, many different alignments for the same sequence set may be equally optimal under the scoring scheme chosen, and it is often difficult to choose among them (Fitch and Smith 1983; Wheeler and Gladstein 1994; Vingron 1996; Cerchio and Tucker 1998; Landan 2005). Furthermore, the accuracy of the alignments obtained using common default values of alignment parameters is only slightly worse than those obtained were we to know the true penalties for inserting gaps and allowing base substitutions (Landan 2005). This means that the absence of knowledge of true parameter value does not substantially affect the alignment accuracy, although it also points to the need for developing methods wherein the true parameters can be used more effectively.

Insurmountable computational demands limit our ability to generate optimal alignment for more than a few sequences. Finding an optimal alignment for a single pair of sequences, even very long genomic ones, has become practical using dynamic pro-

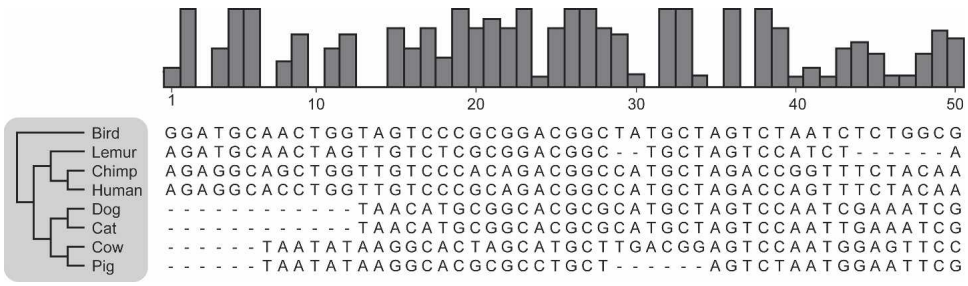

Figure 2. Example of an alignment. A phylogenetic tree is shown for the taxa (rows), and a G+C content profile is shown for the sites (columns).

gramming methods that now require running times and memory requirement proportional to the total lengths of the two sequences (Needleman and Wunsch 1970; Myers and Miller 1988; Delcher et al. 2002; Kurtz et al. 2004). However, the time requirements go up exponentially with increasing numbers of sequences and quickly exceed the capabilities of available computers, even when cleverly optimized (Lipman et al. 1989). Therefore, all practical applications today resort to the use of heuristics when aligning more than a few sequences. Progressive alignment is the most commonly employed heuristic procedure. In this method, a multiple sequence alignment is generated in a step-wise fashion by iteratively aligning pairs of individual sequences or aligned groups of sequences (Barton and Sternberg 1987; Feng and Doolittle 1987; Taylor 1988; Thompson et al. 1994; Brudno et al. 2003a; Bray and Pachter 2004). Because the algorithm needs to start somewhere, a hierarchical alignment order is first created by determining which pairs of sequences and pairs of groups of sequences are the most similar. This order is often referred to as the "guide tree." Using progressive alignment, it was possible to align one thousand short (<500 base pairs each) sequences in a little over an hour using an ordinary PC workstation, and this time was reduced to less than seven minutes using special-purpose hardware (Oliver et al. 1992).

In heuristic approaches, no overall optimal alignment is sought. Instead, it is hoped that optimizing pairwise alignments will lead to a "good" solution. The time efficiency of this approach has led to its becoming the standard, as reflected in its implementation in a variety of well-used software packages (Thompson et al. 1994; Notredame et al. 2000; Brudno et al. 2003a). While some other methods are also available, they have not yet become mainstream because of the lack of unequivocal superiority over classical methods (for review, see Eddy 1995; Notredame and Higgins 1996; Morrison and Ellis 1997; Morgenstern et al. 1998; Thompson et al. 1999; Pollard et al. 2004). Finally, in light of the fact that over 75 approaches to the multiple alignment problem are available (M.S. Rosenberg, pers. comm.), methods have been developed to produce a single consensus alignment out of many (Wallace et al. 2006). This consensus approach remains to be thoroughly evaluated.

Up to this point, we have primarily focused on DNA sequence alignments, but the discussion applies in principle to the alignment of amino acid sequences. In the latter, the probability of substitution from one amino acid to another is incorporated by using 20×20 scoring matrices (e.g., PAM and BLOSUM) to accommodate differences in different types of amino acid substitutions (Dayhoff et al. 1978; Altschul 1991; Henikoff and Henikoff 1992). Furthermore, the alignment of coding DNA sequences of exons is regularly carried out by first aligning the translated amino acid sequences and then adjusting the DNA sequence to reflect the protein alignment. In addition to avoiding the disruption of coding frames, this approach offers the added benefit of achieving better alignment accuracy, because amino acid changes accumulate more slowly than DNA base changes (Zhang et al. 1997; Nei and Kumar 2000).

Finding functionally important motifs via sequence alignment

A major current focus of study in comparative genomics is the identification of short motifs important for gene regulation. Many motif discovery tools have been developed with the common approach to construct a multiple alignment of homologous sequences and identify short stretches of DNA positions that are more conserved over disparate genomes than would be expected by chance (see Stojanovic et al. 1999; Stormo 2000; Keich and Pevzner 2002; McCue et al. 2002; Bulyk 2003; Sinha et al. 2004; Tompa et al. 2005). Accurate motif discovery is known to pose challenging requirements for a multiple alignment program. The data used need to contain sufficiently numerous and diverged species to allow distinction to be made between short significantly conserved (potentially functional) regions and regions that are conserved by chance alone. ENCODE and other projects have addressed this situation for at least some groups of species by providing long homologous DNA segments for many closely related species (particularly mammals) along with some distant relatives of humans (ENCODE Project Consortium 2004; http://www.genome.gov/10005107). The evaluation of the success in finding functionally important segments of DNA (motifs) requires knowledge of the accuracy of current procedures for aligning homologous sites correctly in all the sequences in the data set (perfect columns) as well as the accuracy of aligning successive DNA positions without gaps. Surprisingly, no investigations appear to have tackled both of these questions together in a comprehensive fashion; however, results from some specific studies may be tied together to obtain useful insights.

The accuracy of alignment algorithms has been assessed directly by measuring the fraction of positions that are aligned correctly between sequence pairs in multiple sequence alignments in computer-simulated data (Altschul and Gish 1996; Thompson et al. 1999; Pollard et al. 2004; Rosenberg 2005a). This metric is referred to as "homology accuracy" here. As expected, homology accuracy is found to be lower for distantly related sequences irrespective of the algorithm used (Fig. 3). The usual pattern is that the fraction of positions aligned correctly declines with increasing sequence divergence, with a rapid decline apparent once the difference between DNA sequences is greater than 35% (Fig. 3A). This point corresponds to one base substitution for every two positions, which is close to an evolutionary distance of 0.5 substitutions per site (Pollard et al. 2004; Rosenberg 2005a). This trend holds for several well-known alignment tools; all of them become ineffective by the time the sequence divergence reaches one substitution per site (even though their accuracy diminishes at different rates). To put these divergences in an evolutionary perspective, sequence divergences close to 0.5 substitutions per site are often reported for the comparison of neutrally evolving positions in human and mouse, whereas mouse–chicken and *Drosophila melanogaster–D. pseudoobscura* divergences are expected to exceed one substitution per site (Kumar and Subramanian 2002; Tamura et al. 2004).

Limits on the homology accuracy in Figure 3A are for DNA segments in which all positions are evolving strictly neutrally (i.e., without any natural selection). However, a more realistic scenario is to consider situations in which genomic segments contain highly conserved blocks, because natural selection acts to keep important motifs intact to maintain function. As expected, the existence of conserved blocks enhances the homology accuracy and makes the performance of different methods more similar. Differences do exist, however. Global alignment programs (e.g., ClustalW) perform worse than the procedures that are essentially local in nature (e.g., Lagan, DiAlign), especially for highly divergent sequences (Fig. 3B). The local alignment methods work better because they look for regions of very high sequence similarity and, thus, evolutionary conservation (Smith and Waterman 1981). The presence of large- and small-

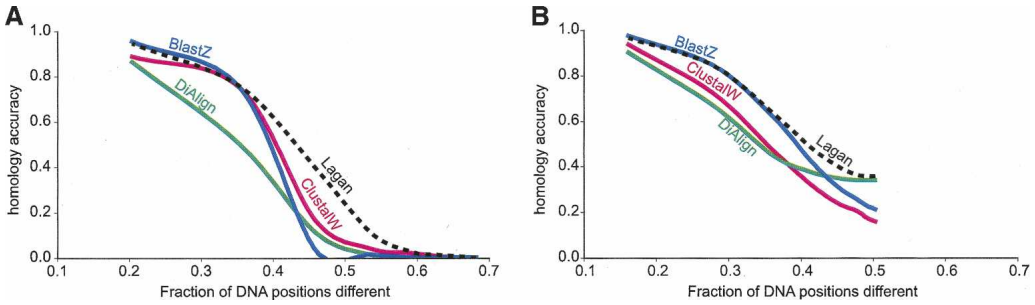

Figure 3. Graphs showing how the accuracy of pairwise alignment varies with evolutionary distance in computer simulations with only insertion–deletions (*A*) and insertion–deletions together with the constraint that 20% of all DNA positions are occupied by interspersed highly conserved blocks that evolve 10 times slower than other positions (Pollard et al. 2004) (*B*). The homology accuracy is calculated as a fraction of simulated positions that were aligned correctly, and it is plotted against the fraction of sites different, as reported in Figures 2 and 4 of Pollard et al. (2004). Even though the same set of evolutionary distances is simulated for both panels, the percent sequence difference in *B* is smaller for a given simulated distance because of the existence of highly conserved blocks. Programs compared for homology accuracy are ClustalW (Thompson et al. 1994), BLASTZ (Schwartz et al. 2003), DiAlign (Morgenstern 1999), and Lagan (Brudno et al. 2003a).

scale rearrangements will also severely affect the quality of the global alignments (Morgenstern et al. 1998; Siddharthan 2006).

The probability of aligning multiple adjacent sites correctly in a set of sequences is a direct function of the homology accuracy at each position, so it is evident that identifying short genomic segments involved in gene regulation via comparative sequence analysis is likely to produce many false-negatives if the sequences involved are highly divergent. This is clearly seen in a dramatic decline in finding short motifs after the addition of a nonmammalian species (chicken) to a data set containing placental mammals (human, chimp, mouse, and rat) (Prakash and Tompa 2005) (Fig. 4). Because the bird–mammal sequence divergence is expected to be about three times the divergence between human and mouse based on the timing of their divergence (Hedges and Kumar 2003), the inclusion of an avian sequence demands that the alignment methods correctly align DNA bases that have experienced up to 1.5 mutations per site. The homology accuracy is very low at such high divergences, as mentioned above, and the chances of correctly aligning conserved motifs are small (Fig. 3). Of course, it is possible, and would not be surprising, that fewer conserved motifs are found between birds and mammals simply because fewer such motifs exist (Cooper et al. 2005).

The use of multiple sequences and the choice of species critically impacts motif discovery. Certainly, pairwise sequence analysis can be carried out in the absence of a robust phylogeny (Elnitski et al. 2003), but this is less informative because of the difficulty in detecting multiple substitutions at the same site, and because of an inability to distinguish the direction of sequence change (Bergman and Kreitman 2001). Also, the use of multiple sequences improves statistical power to detect short conserved elements by bringing more data to bear on the problem (Sumiyama et al. 2001; Gottgens et al. 2002; Brudno et al. 2003a; Thomas et al. 2003). The term "data" here refers not only to the sequence length and the total number of differences among sequences but also to the distribution of sequence differences among the species represented. For example, a set of several closely related (e.g., primate) sequences tends to be more informative than a single pair of sequences (e.g., mouse and human), even though the total number of substitutions in the evolutionary history of the sequences considered is identical. We expect better homology accuracy in primate sequence alignments than that between human and mouse sequences, so we expect to find more highly conserved motifs when using the primate data set. This was indeed the case when a group of primate species was used to provide essentially the same amount of collective additive phylogenetic diver-

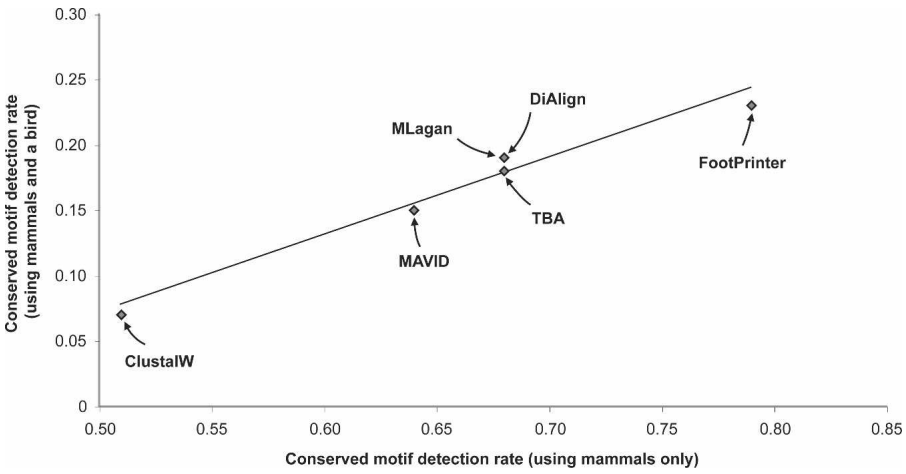

Figure 4. Graph showing the performance of different multiple sequence alignment programs in detecting conserved motifs of length 10 in sets of orthologous 1000 base-pair putative promoter regions of vertebrate genomes (Prakash and Tompa 2005). The axes represent the fraction of alignments (using the given program) of orthologous sequences in which at least one perfectly conserved motif of length 10 was detected. The *x*-axis represents 5073 alignments containing human, chimp, mouse, and rat sequences, while the *y*-axis represents 945 alignments of human, chimp, mouse, rat, and chicken. The smaller number of hits in the latter set is partly attributable to fewer conserved motifs and partly to increased difficulty of detection. The programs used are ClustalW (Thompson et al. 1994), MAVID (Bray and Pachter 2004), DiAlign (Morgenstern 1999), MLagan (Brudno et al. 2003a), TBA (Blanchette et al. 2004), and FootPrinter (Blanchette and Tompa 2003). Data are from Figure 2 of Prakash and Tompa (2005).

gence as exists between mouse and human (Boffelli et al. 2003).

Many DNA segments will be found to be conserved in closely related species, but such false positives can be detected by using many sequences and by estimating false-positive rates (Cooper et al. 2005; Prakash and Tompa 2005). Still, all conserved motifs found may not be functional, and the pursuit of functionally important motifs via evolutionary conservation profiles implicitly assumes that the same motif(s) are functional in different species. The latter assumption is expected to be satisfied in analyses involving closely related species as compared with distantly related species (Cooper et al. 2005). Therefore, the species chosen are critical in multispecies analysis to find candidate, functional motifs.

Knowledge of correct evolutionary relationships is important in motif discovery. Efforts have been made to determine how the use of shared ancestry (specified using phylogenetic relationships) enhances the accuracy of motif detection over simple treatment of sequences as an aligned set (Dubchak and Frazer 2003; Sinha et al. 2004; Siddharthan et al. 2005; Gertz et al. 2006). However, the effect of using incorrect phylogeny on the accuracy of motif-discovery remains to be measured. Intuitively, it is clear that if two sister species are placed distantly in a phylogeny, then statistical methods will spuriously identify conserved regions as candidate motifs, even though they could be explained by chance alone if the correct phylogeny were to be used. Likewise, placing species nearer than they ought to be would discount the significance of similarities and lead to false-negatives. There is an urgent need for quantification of the false-positive and -negative rates in motif discovery caused by the use of incorrect phylogenetic trees.

Estimating sequence divergence

Evolutionary distances are routinely estimated from pairs of aligned sequences and are used for inferring phylogenies, divergence times, and rates of evolution (Nei and Kumar 2000). What effect does the alignment process have on distances estimated in this way? It appears that as long as the evolutionary distance is less than about one substitution per site between sequences, evolutionary distances estimated from pairwise alignments of DNA sequences are relatively insensitive to even large amounts (50%) of alignment error (Rosenberg 2005a). In contrast, and as expected, computer simulations exploring the effect of the use of a wide range of alignment parameter values leads to distances that often fall outside the 95% confidence interval around the optimal estimates (Fleissner et al. 2000, 2005) (Fig. 5), because a large fraction of alignment parameter values are likely to be biologically unrealistic. There is no single default set that works over a wide range of true distances and insertion–deletion models, even though methods are available to help select good values for alignment parameters (e.g., Holmes and Durbin 1998).

Multiple sequence alignments are considered better than pairwise alignments because more similar sequences will act as intermediates between highly dissimilar ones (Lesk 2005). How does adding a third sequence to an alignment problem affect the accuracy of subsequently estimated distances? In a direct two-sequence comparison, a branch to a third sequence was added at varying intermediate points between the two. This improved the accuracy of estimated distance between the original two, but the highest accuracy for the estimates of evolutionary distance was not achieved when the maximal homology accuracy was found.

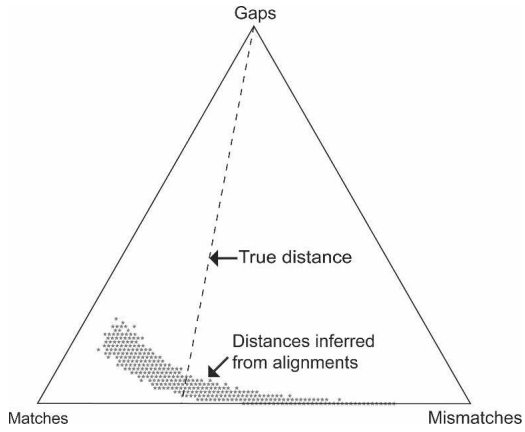

Figure 5. A barycentric representation of the distribution of matches, mismatches, and gaps for optimally aligned pairs of sequences having exactly 60 matches, 30 mismatches, and 10 gaps each over the entire range of alignment parameters (match, mismatch, and gap penalties) for which mismatches are penalized more than matches. Each point in the triangle represents a number of matches, mismatches, and gaps summing to 100. The lengths of the perpendiculars from a point in the diagram to the *right*, *left*, and *bottom* side of the triangle are proportional to the corresponding match, mismatch, and gap fractions, respectively. Simulation outcomes are represented by the dark gray points near the *bottom* of the triangle. In 1000 trials, the true sequence-pair description (60 matches, 30 mismatches, 10 gaps) was never recovered, and only those 1.1% of the alignments lying precisely on the line labeled "true distance" yielded estimated distances equal to the true distance (1/3) between the original pair of sequences. Adapted, with permission, from Figure 2.3 of Fleissner 2003.

Generally, the estimated distance between the two original sequences increased fairly linearly as the origin of the added branch was moved closer to the common ancestor of the original pair and the true evolutionary divergence was less than one substitution per site (Rosenberg 2005b).

Another measure of homology accuracy useful for motif detection is the fraction of sites successfully aligned when sequences are added one-by-one to the data set. This is used when the true sequence alignment is not known, as in empirical data analysis. Indeed, the number of sites aligned between two distantly related sequences increases as the data set is expanded by adding more sequences intermediate to two distantly related sequences in question (Margulies et al. 2006). This is similar to the effect observed for homology accuracy in the simulated data (Rosenberg 2005b), with the caveat that the relative divergence of the additional sequence is critically important.

Effect of alignment homology inaccuracies on phylogenetic inference

The common modus operandi for building phylogenetic trees is to align a data set using some program, inspect the result for obvious error, perhaps realign, and then infer a phylogeny from the result. Current computer simulations aimed at deciphering the effect of sequence homology accuracy on phylogenetic reconstruction for various shapes of trees and phylogenetic methods show that, on average, homology accuracy correlates with the accuracy of the inferred phylogeny (Ogden and Rosenberg 2006). However, the relationship is far from monotonic and phylogenetic accuracy may increase or decrease dramatically even with small changes in homology accuracy (Fig. 6). In fact, the phylogenies inferred appear to be more sensitive to alignment method and parameters than to the choice of the tree-building

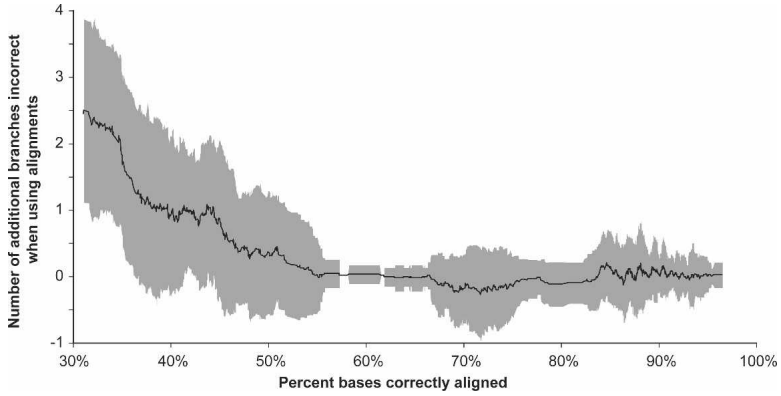

Figure 6. Graph illustrating the relationship between the accuracy of an alignment and the topological error of a phylogenetic tree reconstructed from that alignment by the Maximum Likelihood method in a large-scale simulation study. The *x*-axis is the average homology accuracy of all pairwise alignments in a multiple alignment. The *y*-axis is the number of incorrect branches in the tree obtained from the reconstructed alignment minus the number of incorrect branches in the tree reconstructed from the true alignment. Each point plotted is a moving average of 50 simulated results. Trees are based on 16 taxa and have a variety of topologies, relative branch lengths, and maximum distances. The shaded area shows plus and minus one standard deviation of points within the moving-average window. The results indicate that very poor alignments produce bad trees, but as long as 60% or more of the sites are accurately aligned, further improvements in alignment homology accuracy make little difference; even highly accurate alignments produce trees with substantial variation in quality. Results for Bayesian and Maximum Parsimony phylogenetic analyses are similar; Neighbor-Joining shows higher levels (an average of around 1 additional branch incorrect) of phylogenetic error. Based on data from Figure 4 of Ogden and Rosenberg (2006).

method (Morrison and Ellis 1997; Cammarano et al. 1999; Ogden and Whiting 2003; Hertwig et al. 2004; Lebrun et al. 2006).

The inevitable need for the use of heuristic procedures in sequence alignment contributes significantly toward phylogenetic error, when the sequences being aligned have undergone a large number of substitutions and many insertion–deletion events. A key component of any progressive (heuristic) alignment procedure is the guide tree that sets up the order in which sequences (and sequence profiles) are aligned. Guide trees can be created in different ways. ClustalW, for example, creates a matrix of pairwise distances by aligning each sequence pair separately and computing a dissimilarity score from each pairwise alignment. This dissimilarity matrix is then subjected to the neighbor-joining algorithm to generate a directional hierarchy of sequence relationships (Thompson et al. 1994). Obviously, the guide tree is not guaranteed to reflect the true evolutionary history of sequences; because the dissimilarity scores are not evolutionary distances, the sequence alignment produced will most likely be based on incorrect inferences.

Guide-tree errors are known to have serious effects on downstream phylogenetic inference, as the increase in the phylogenetic error rate is found to be associated with errors in the guide trees (Lake 1991; Thorne and Kishino 1992; Landan 2005; Redelings and Suchard 2005). However, the error introduced by incorrect guide trees is generally considered a nuisance and thought to decrease the power of statistical inference by reducing our ability to distinguish among alternative phylogenetic hypotheses. Because the use of a guide tree is required to make sequence alignment feasible computationally, the implicit consolation has been that at least incorrect phylogenetic clusters will not garner high statistical support. This intuition is reasonable for short sequences, as the statistical biases introduced by guide trees in sequence alignments are expected to be smaller than the variances associated with estimation of evolutionary distances and branch lengths when conducting Maximum Likelihood and Least Squares analyses using sophisticated models of nucleotide substitution.

Disregarding the effect of guide-tree errors on evolutionary and phylogenetic inference is no longer tenable, because these errors are amplified in today's large data sets. The genomic revolution in building the Tree of Life has now taken root and very long sequences are being used to establish key species relationships (Hedges 2002; Bashir et al. 2005; Margulies et al. 2005; Elango et al. 2006). First, these sequences need to be aligned, which is done invariably by using guide trees either explicitly or implicitly. In such data sets, the variance of estimates of evolutionary parameters is expected to become vanishingly small, as it decreases with sequence length. Alignment bias caused by errors in the guide tree will affect each position aligned, and it will not diminish with increasing sequence length. These two factors may combine to mislead phylogenomic inferences and incorrect phylogenetic clusters may be supported by spuriously high confidence. For example, alignments of homologous 20,000 base-pair regions of human, mouse, rabbit, and cow produce high bootstrap support for the species relationship implied by the guide tree itself when ClustalW is used (Fig. 7). A similar trend is observed when using genome sequence alignment approaches, such as the TBA (Blanchette et al. 2004) (V. Swarna, pers. comm.). We do not know which phylogeny is the true one, but it is clear that error in guide trees can produce incorrect inferences supported by high confidence. On a positive note, guide trees inferred using very long sequences may contain few or no errors, so the situation may not be as grim as it appears. However, this question has yet to be examined and is being investigated currently by our group.

When is it appropriate to use genomic multiple sequence alignments available in various database resources, which use a specific phylogeny as a guide tree? The answer depends on the purpose of the analysis. To begin with, the use of such multiple sequence alignments for inferring species phylogenies will be circular and will often (but not always) produce outcomes that merely reflect bias introduced by the alignment procedure (Fig. 7). This will be particularly problematic for traditionally hard-to-resolve phylogenetic relationships, because they are often associated with short internal branches (i.e., small amount of evolutionary change) in the phylogenetic trees. Errors introduced by alignment bias are expected to affect these parts of the phylogeny most severely, as the phylogenetic signal may be overwhelmed by the bias in the sequence alignment. For example, resolving the phylogenetic relationship of major groups of mammals using the sequence alignment of ENCODE data is not desirable, because these alignments are constructed using a guide tree that best reflects our current understanding of mammalian and vertebrate species relationships (http://www.genome.gov/10005107). On the other hand, these alignments are appropriate for inferring ancestral genomes and times of species divergence events, because

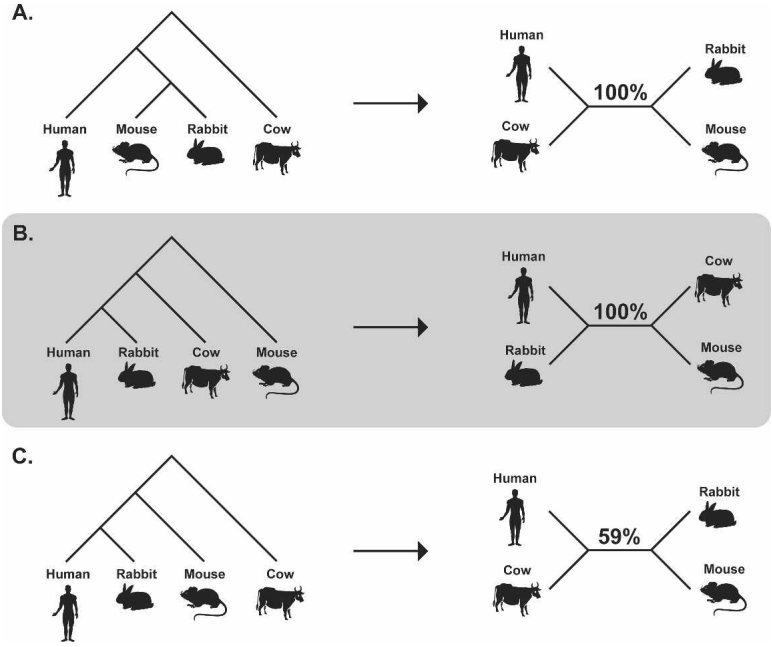

Figure 7. Diagrams show different guide trees used in a progressive alignment (*left*) and the phylogenies inferred from the resulting alignments (*right*). The bootstrap support for the inferred interior branch is shown (1000 replicates). Three different rooted guide trees (*A*, *B*, and *C*) were used to align 20,000 base pairs from human, mouse, rabbit, and cow sequences in the ENCODE data set ENm001 (ENCODE Project Consortium 2004; http://www.genome.gov/10005107). Default options in ClustalW (Thompson et al. 1994) were used and the phylogenetic trees were inferred from the alignments using the Neighbor-Joining (NJ) method as implemented in MEGA3 (Kumar et al. 2004). (Maximum Likelihood (ML) analyses also support the same phylogenies.) All sites containing alignment gaps were removed before phylogenetic analysis, resulting in a data set of more than 8000 positions in each alignment. Because the NJ and ML methods both produce unrooted phylogenies, results are shown as such. In *A* and *B*, the contradictory inferred phylogenetic trees are identical to the guide trees, and each obtains 100% NJ bootstrap support. The influence of the root of the guide tree can be seen by comparing results in *B* and *C*: The guide trees are identical except for the location of the root, yet the alignments they produce result in incompatible inferences about the relationships of species.

those procedures and all the results obtained are explicitly conditional on the evolutionary tree used. Obviously, one would generally use the same evolutionary tree for aligning sequences and for conducting evolutionary analyses, as long as it is substantiated. Otherwise, the use of an incorrect phylogeny will not only produce ancestral sequences and speciation times for ancestors that never existed, but it will also bias results for those that indeed existed. On the other hand, errors in the guide tree are expected to have a significantly lesser impact on the estimation of evolutionary rates at individual positions (or in sliding windows), and thus on motif finding, as such summary statistics are derived using a large amount of data for each position when many species are used (see, e.g., Yang and Kumar 1996; Siepel et al. 2005).

One unexplored wrinkle in the guide-tree conundrum is the observation that closely related species can show very different base compositions in homologous genomic segments. For instance, equality of substitution pattern can be rejected in ~40% of human–mouse protein-coding gene orthologs, associated with differences in G+C content (Kumar and Gadagkar 2001; Jermiin et al. 2004). Among eubacteria, genome-wide G+C content can vary by a factor of two (Nakashima et al. 2003). This variation would serve to distort apparent distances among taxa and lead to biased guide trees. Either one may bias phylogenetic analyses (Steel et al. 1993; Kolaczkowski and Thornton 2004; Gadagkar and Kumar 2005). These deviations are seldom incorporated in simulated studies or models, but they are known to run rampant in real sequence data. One way of avoiding guide-tree bias on phylogenetic inference is to simultaneously infer multiple sequence alignment and phylogeny, so that one does not depend on the completion of the other (Hein 1990). Maximum Likelihood and Bayesian methods developed for this purpose seem promising for small-to-medium-sized problems (Fleissner et al. 2005; Lunter et al. 2005; Redelings and Suchard 2005), but their application to data sets with a large number of sequences and to long sequences is still computationally prohibitive.

Conclusions

The multiple sequence alignment procedure forms the backbone of comparative and evolutionary genomics. Results from some recent studies involving computer simulations and large-scale genomic data have begun to clarify and quantify sources of bias and the effects of alignment on subsequent downstream processing. Because of rapid growth of large scale data sets and increasing applications of multiple sequence alignments to understand patterns and processes that govern gene, genome, and species evolution, it would be prudent to further intensify these investigations and make their conclusions more accessible to practicing biologists.

Acknowledgments

We thank Vinod Swarna for his assistance with data analysis (Fig. 7), Drs. Sonja Prohaska and Michael Rosenberg for comments on an earlier version of this manuscript, and Ms. Kristi Garboushian for editorial support. We also thank three anonymous referees for many insightful suggestions. This work was supported in part by a research grant from National Institutes of Health to S.K.

References

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219:** 555–565.
Altschul, S.F. and Gish, W. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.
Barton, G.J. and Sternberg, M.J. 1987. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198:** 327–337.
Bashir, A., Ye, C., Price, A.L., and Bafna, V. 2005. Orthologous repeats and mammalian phylogenetic inference. *Genome Res.* **15:** 998–1006.
Bergman, C.M. and Kreitman, M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11:** 1335–1345.
Blanchette, M. and Tompa, M. 2003. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.* **31:** 3840–3842.
Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.
Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I.,

Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.
Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14:** 693–699.
Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003a. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721–731.
Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., and Batzoglou, S. 2003b. Glocal alignment: Finding rearrangements during alignment. *Bioinformatics* **19:** i54–i62.
Bulyk, M.L. 2003. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5:** 201.
Cammarano, P., Creti, R., Sanangelantoni, A.M., and Palm, P. 1999. The archaea monophyly issue: A phylogeny of translational elongation factor G(2) sequences inferred from an optimized selection of alignment positions. *J. Mol. Evol.* **49:** 524–537.
Cerchio, S. and Tucker, P. 1998. Influence of alignment on the mtDNA phylogeny of Cetacea: Questionable support for a Mysticeti/Physeteroidea clade. *Syst. Biol.* **47:** 336–344.
Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302:** 1960–1963.
Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15:** 901–913.
Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.Ø. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92–96.
Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (ed. M.O. Dayhoff), pp. 345–352. National Biomedical Research Foundation, Washington, D.C.
Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30:** 2478–2483.
Dubchak, I. and Frazer, K. 2003. Multi-species sequence comparison: The next frontier in genome annotation. *Genome Biol.* **4:** 122.
Eddy, S.R. 1995. Multiple alignment using hidden Markov models. In *Proceedings of the Third International Conference on Intelligence Systems for Molecular Biology*, pp. 114–120. AAAI Press, Menlo Park, CA.
Elango, N., Thomas, J.W., and Yi, S.V. 2006. Variable molecular clocks in hominoids. *Proc. Natl. Acad. Sci.* **103:** 1370–1375.
Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W., and Chiaromonte, F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64–72.
ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.
Felsenstein, J. 2002. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.
Feng, D.F. and Doolittle, R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25:** 351–360.
Fitch, W.M. and Smith, T.F. 1983. Optimal sequence alignments. *Proc. Natl. Acad. Sci.* **80:** 1382–1386.
Fleissner, R. 2003. "Sequence Alignment and Phylogenetic Inference." Ph.D. thesis, Heinrich-Heine-Universität, Düsseldorf.
Fleissner, R., Metzler, D., and von Haeseler, A. 2000. Can one estimate distances from pairwise sequence alignments? In *German Conference on Bioinformatics* (eds E. Bornberg-Bauer et al.), pp. 89–96. Logos Verlag, Heidelberg.
Fleissner, R., Metzler, D., and von Haeseler, A. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol.* **54:** 548–561.
Gadagkar, S.R. and Kumar, S. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Mol. Biol. Evol.* **22:** 2139–2141.
Gaucher, E.A., Gu, X., Miyamoto, M.M., and Benner, S.A. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* **27:** 315–321.
Gertz, J., Fay, J.C., and Cohen, B.A. 2006. Phylogeny based discovery of regulatory elements. *BMC Bioinformatics* **7:** 266.
Gottgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)–comparative analysis of five vertebrate SCL loci. *Genome Res.* **12:** 749–759.
Gu, X. and Li, W.H. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J. Mol. Evol.* **40:** 464–473.
Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.
Hedges, S.B. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3:** 838–849.
Hedges, S.B. and Kumar, S. 2003. Genomic clocks and evolutionary timescales. *Trends Genet.* **19:** 200–206.
Hein, J. 1990. Unified approach to alignment and phylogenies. *Methods Enzymol.* **183:** 626–645.
Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.
Hertwig, S., de Sá, R.O., and Haas, A. 2004. Phylogenetic signal and the utility of 12S and 16S mtDNA in frog phylogeny. *J. Zoological Syst. Evol. Res.* **42:** 2–18.
Holmes, I. and Durbin, R. 1998. Dynamic programming alignment accuracy. *J. Comput. Biol.* **5:** 493–504.
Jermiin, L., Ho, S.Y., Ababneh, F., Robinson, J., and Larkum, A.W. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol.* **53:** 638–643.
Keich, U. and Pevzner, P.A. 2002. Subtle motifs: Defining the limits of motif finding algorithms. *Bioinformatics* **18:** 1382–1390.
Kolaczkowski, B. and Thornton, J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431:** 980–984.
Koonin, E.V. and Galperin, M.Y. 2003. *Sequence—Evolution—Function: Computational Approaches in Comparative Genomics*. Kluwer Academic, Boston.
Kumar, S. and Gadagkar, S.R. 2001. Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* **158:** 1321–1327.
Kumar, S. and Subramanian, S. 2002. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci.* **99:** 803–808.
Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief. Bioinform.* **5:** 150–163.
Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* **5:** R12.
Lake, J.A. 1991. The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.* **8:** 378–385.
Landan, G. 2005. Multiple sequence alignment errors and phylogenetic reconstruction. In *Zoology*, pp. 93. Tel Aviv University, Tel Aviv.
Lebrun, E., Santini, J.M., Brugna, M., Ducluzeau, A.L., Ouchane, S., Schoepp-Cothenet, B., Baymann, F., and Nitschke, W. 2006. The Rieske protein: A case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction. *Mol. Biol. Evol.* **23:** 1180–1191.
Lesk, A.M. 2005. *Introduction to bioinformatics*. Oxford University Press, Oxford, New York.
Lipman, D.J., Altschul, S.F., and Kececioglu, J.D. 1989. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci.* **86:** 4412–4415.
Lunter, G., Miklos, I., Drummond, A., Jensen, J.L., and Hein, J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6:** 83.
Margulies, E.H., Vinson, J.P., Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., and Clamp, M. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci.* **102:** 4795–4800.
Margulies, E.H., Chen, C.W., and Green, E.D. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* **22:** 187–193.
McCue, L.A., Thompson, W., Carmack, C.S., and Lawrence, C.E. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* **12:** 1523–1532.
Miklos, I., Lunter, G.A., and Holmes, I. 2004. A "Long Indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.* **21:** 529–540.
Morgenstern, B. 1999. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15:** 211–218.
Morgenstern, B., Frech, K., Dress, A., and Werner, T. 1998. DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics* **14:** 290–294.
Morgenstern, B., Prohaska, S.J., Pohler, D., and Stadler, P.F. 2006. Multiple sequence alignment with user-defined anchor points. *Algorithms Mol. Biol.* **1:** 6.
Morrison, D.A. and Ellis, J.T. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of apicomplexa. *Mol. Biol. Evol.* **14:** 428–441.
Myers, E.W. and Miller, W. 1988. Optimal alignments in linear space.

Comput. Appl. Biosci. **4:** 11–17.
Nakashima, H., Fukuchi, S., and Nishikawa, K. 2003. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J. Biochem.* **133:** 507–513.
Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.
Nei, M. and Kumar, S. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, Oxford, UK.
Notredame, C. and Higgins, D.G. 1996. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res.* **24:** 1515–1524.
Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.
Ogden, T.H. and Whiting, M.F. 2003. The problem with the "Paleoptera Problem:" Sense and sensitivity. *Cladistics* **19:** 432–444.
Ogden, T.H. and Rosenberg, M.S. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* **55:** 314–328.
Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P., Benit, P., et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38–46.
Pevsner, J. 2003. *Bioinformatics and functional genomics*. Wiley-Liss, Hoboken, NJ.
Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., and Eisen, M.B. 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5:** 6.
Prakash, A. and Tompa, M. 2005. Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* **23:** 1249–1256.
Qian, B. and Goldstein, R.A. 2001. Distribution of indel lengths. *Proteins* **45:** 102–104.
Redelings, B.D. and Suchard, M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* **54:** 401–418.
Rosenberg, M.S. 2005a. Evolutionary distance estimation and fidelity of pair wise sequence alignment. *BMC Bioinformatics* **6:** 102.
Rosenberg, M.S. 2005b. Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinformatics* **6:** 278.
Rosenberg, M.S. and Kumar, S. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Mol. Biol. Evol.* **20:** 610–621.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Siddharthan, R. 2006. Sigma: Multiple alignment of weakly-conserved non-coding DNA sequence. *BMC Bioinformatics* **7:** 143.
Siddharthan, R., Siggia, E.D., and van Nimwegen, E. 2005. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* **1:** e67.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Sinha, S., Blanchette, M., and Tompa, M. 2004. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5:** 170.
Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.
Steel, M.A., Lockhart, P.J., and Penny, D. 1993. Confidence in evolutionary trees from biological sequence data. *Nature* **364:** 440–442.
Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W., and Hardison, R. 1999. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.* **27:** 3899–3910.
Stormo, G.D. 2000. DNA binding sites: Representation and discovery. *Bioinformatics* **16:** 16–23.
Sumiyama, K., Kim, C.B., and Ruddle, F.H. 2001. An efficient *cis*-element discovery method using multiple sequence comparisons based on evolutionary relationships. *Genomics* **71:** 260–262.
Tamura, K., Subramanian, S., and Kumar, S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21:** 36–44.
Taylor, W.R. 1988. A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* **28:** 161–169.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustal-W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27:** 2682–2690.
Thorne, J.L. and Kishino, H. 1992. Freeing phylogenies from artifacts of alignment. *Mol. Biol. Evol.* **9:** 1148–1162.
Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23:** 137–144.
Town, C. 2002. *Functional genomics*. Kluwer Academic, Boston, MA.
Vingron, M. 1996. Near-optimal sequence alignment. *Curr. Opin. Struct. Biol.* **6:** 346–352.
Vingron, M. and Waterman, M.S. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* **235:** 1–12.
Vogel, F. and Kopun, M. 1977. Higher frequencies of transitions among point mutations. *J. Mol. Evol.* **9:** 159–180.
Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. 2006. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34:** 1692–1699.
Waterman, M.S., Smith, T.F., and Beyer, W.A. 1976. Some biological sequence metrics. *Adv. Math.* **20:** 367–387.
Wheeler, W.C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Syst. Biol.* **44:** 321–331.
Wheeler, W.C. and Gladstein, D.S. 1994. MALIGN: A multiple sequence alignment program. *J. Hered.* **85:** 417–421.
Yang, Z. and Kumar, S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **13:** 650–659.
Yuan, J., Amend, A., Borkowski, J., DeMarco, R., Bailey, W., Liu, Y., Xie, G., and Blevins, R. 1999. MULTICLUSTAL: A systematic method for surveying Clustal W alignment parameters. *Bioinformatics* **15:** 862–863.
Zhang, Z., Pearson, W.R., and Miller, W. 1997. Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.* **4:** 339–349.
Zhu, J., Liu, J.S., and Lawrence, C.E. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14:** 25–39.