Science
AAAS

Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis

Ari Löytynoja, *et al.*

Science **320**, 1632 (2008);
DOI: 10.1126/science.1158395

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2008 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis

Ari Löytynoja* and Nick Goldman

Genetic sequence alignment is the basis of many evolutionary and comparative studies, and errors in alignments lead to errors in the interpretation of evolutionary information in genomes. Traditional multiple sequence alignment methods disregard the phylogenetic implications of gap patterns that they create and infer systematically biased alignments with excess deletions and substitutions, too few insertions, and implausible insertion-deletion–event histories. We present a method that prevents these systematic errors by recognizing insertions and deletions as distinct evolutionary events. We show theoretically and practically that this improves the quality of sequence alignments and downstream analyses over a wide range of realistic alignment problems. These results suggest that insertions and sequence turnover are more common than is currently thought and challenge the conventional picture of sequence evolution and mechanisms of functional and structural changes.

New DNA sequencing methods permit quick and affordable exploration of genomic sequences of different organisms. Some of the greatest beneficiaries of the rapid increase of sequence data are comparative genomic studies that seek to provide increasingly accurate reconstruction of evolutionary histories of related genomes, e.g., to study functional and structural sequence changes leading to phenotypic differences between species (*1*–*4*). However, all sequence analyses that rely on evolutionary information require an accurate sequence alignment, i.e., the correct identification of homologous nucleotides or amino acids and the positioning of gaps indicating inserted and deleted sequence.

Alignment is still a highly error-prone step in comparative sequence analysis. Different multiple sequence alignment methods often lead to drastically different conclusions in both phylogenetic analyses and functional studies (supporting online material text), and alternative alignments of the same data can support entirely different mechanisms driving evolutionary and functional changes in sequences. As an example, a traditional alignment of HIV and SIV envelope glycoprotein gp120 (*5*) (Fig. 1A) has a familiar pattern of insertions and deletions squeezed compactly between conserved blocks of structurally important residues and suggests that part of the variable V2 region has a high amino acid–substitution rate and has shortened over time at a mutation hotspot where overlapping sites have been independently deleted in different evolutionary branches: some sites as many as eight times among the 23 sequences included. With an alignment method that considers the sequences' phylogeny and distinguishes insertions from deletions (*5*), the story is different: Instead of multiple point substitutions and loss of sequence, the region evolves through short insertions and deletions, allowing for rapid and radical changes in the coding sequence (Fig. 1B). The latter alignment, which suggests rapid turnover of sequence material instead of long ancestral sequences shrinking in length, provides a more convincing mechanism for the evolution of this region. Furthermore, its association of gap patterns with meaningful insertion and deletion events at the branches of the phylogenetic tree, i.e., specific points in the history of the sequences, allows a realistic reconstruction of the evolutionary process leading to the present-day sequences. In this example, the different implications of the alternative alignments for the mechanisms and time scale of sequence changes may be of medical importance for understanding the evolutionary dynamics of HIV (*6*), particularly in this protein region where insertions, deletions, and substitutions are associated with the efficiency of HIV entry, biological phenotype, and neutralizing antibody response (*7*–*11*).

Progressive algorithms (*12*–*15*), the multiple sequence alignment methods most widely used today, are based on backtracking the evolutionary process and building a multiple alignment from pairwise alignments between sequences and sequence alignments, performed in order of decreasing relatedness (Fig. 2) (supporting online material text). However, whereas insertion and deletion events are indistinguishable when comparing one pair of sequences, the two events differ greatly in progressive iteration of pairwise alignments. A gap for a deletion, with its associated penalty, is created only once, but a gap for an insertion has to be opened multiple times (Fig. 2, A and B). Simple iteration associates a full penalty with each of these gap-opening events, which leads to excessive penalization of single insertion events.

No alignment methods have previously implemented a precise solution to this problem; instead, heuristics to lower the penalty for opening gaps at positions already containing gaps have been used (*12*, *14*). Although these site-specific penalties reduce the high overall cost of single insertion events and encourage subsequent alignment iterations to correctly place their gaps at the same position, the approach fails when there are multiple nearby insertions and deletions and becomes systematically biased. By definition, inserted characters are not descendants of—and thus are not homologous with—any other insertions or ancestral characters, and should never align with anything (Fig. 2C, evolution). Progressive algorithms, however, always incorrectly align neighboring insertions in the same column if that is not explicitly prevented; the use of site-specific gap penalties, instead of preventing the incorrect matching of independent insertions, encourages it (Fig. 2C, site-specific alignment). Such "collapsed insertions" create incorrect homologies and, as the resulting gap pattern implies multiple independent deletions, give an impression of deletion hotspots where the overly long ancestral sequences are shortened (Fig. 2C, interpretation). In addition, the procedure also lowers the penalties at deletion sites where no further gaps are required, creating "gap magnets" that make nearby deletions coincide in subsequent stages of progressive iteration (Fig. 2D, evolution and site-specific alignment). Similarly to incorrectly aligned insertions, the clustering of deletions creates false homologies and gives an impression of deletion hotspots (Fig. 2D, interpretation).

We previously identified the problem of multiple penalization of insertions and reported a preliminary attempt to solve it (*16*). This uses a phylogeny-aware approach that "flags" the gaps made in previous alignments and, using evolutionary information from related sequences to indicate whether each gap has been created by an insertion or a deletion, permits their "reuse" for inserted characters without further penalty in the next stage of the progressive alignment (Fig. 2C, phylogeny-aware alignment). In addition, information from closely related sequences can be used to infer sites as "permanent" insertions that cannot be matched in subsequent alignments (*5*), so that distinct insertion events are correctly kept separate even when they occur at exactly the same position. If related sequences indicate that a gap is caused by a deletion, flags are removed and no further free gaps at that position are permitted (Fig. 2D), and the effect is correctly targeted on insertions only.

To understand the type and magnitude of algorithm-based errors in traditional sequence

European Molecular Biology Laboratory—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK.

*To whom correspondence should be addressed. E-mail: ari@ebi.ac.uk

alignment methods, we compared the accuracy of different variants of the progressive algorithm, including our implementation of the new phylogeny-aware algorithm distinguishing insertions and deletions (as described above). We simulated synthetic DNA sequence data according to 16-, 32-, and 64-taxon symmetric trees using realistic evolutionary parameters, mimicking the evolution of genomic DNA without the structural and functional constraints expected in protein-coding regions and so that the true alignments contained equal numbers of insertions and deletions (*5*). For the 16-taxon tree, we set evolutionary relationships close, intermediate, and distant (Fig. 3, see color gradients), approximately representing comparisons of primates, primates and rodents, and mammals, respectively. Using the 32- and 64-taxon trees and the maximum species divergence of the close set, we assessed the effects of denser sampling (2X and 4X, respectively) of increasingly similar sequences (fig. S1). The sequences were aligned by using a set of published alignment software programs based on variants of the traditional progressive algorithm [CLUSTAL W (*12*), MAFFT (*15*), MUSCLE (*14*) and T-COFFEE (*13*)] and the phylogeny-aware algorithm [PRANK (*16*); we used the PRANK$_{+F}$ variant indicating "permanent" insertions (*5*)]. For each alignment, various statistics describing the inferred insertion-deletion processes and the accuracy of the solution were computed.

The alignments generated by the alternative methods vary greatly even for the closely related sequences. The methods implementing the traditional algorithm produce alignments with all the errors expected from a biased, nonphylogenetic handling of insertions and deletions. The failure to separate distinct, nearby insertions leads to underestimation of their true number (Fig. 3A) and overestimation of the number of deletions (Fig. 3B); this gives seriously incorrect estimates of the insertion rate/deletion rate ratio (Fig. 3C). Collapsed insertions and gap magnets create an impression of mutation hotspots where the same sequence sites are deleted multiple times [indicated by the "gap overlap" statistic (Fig. 3D)]. These problems make the alignments overly compact and are reflected in the proportion of alignment columns recovered entirely correctly (Fig. 3, E and F). In contrast to the traditional algorithm, the phylogeny-aware PRANK$_{+F}$ program is not systematically biased. It has slightly superior performance in terms of alignment length and proportion of correct columns but, crucially, it is unbiased with respect to insertions and deletions and has virtually no error in all the other measures of insertion and deletion parameters.

As distances between sequences increase, the greater numbers of insertions, deletions, and substitutions make the sequences more difficult to align. The proportion of columns correct is a very stringent measure of alignment accuracy; even under this measure, PRANK$_{+F}$ clearly performs best in response to increased sequence divergence (Fig. 3F, close-intermediate-distant). However, this masks much deeper underlying problems in the traditional algorithms, as can be clearly seen in the growth of the errors in all statistics describing the insertion-deletion processes (Fig. 3, A to E, close-intermediate-distant). In contrast, PRANK$_{+F}$, already more accurate for close sequences, shows superior performance as evolutionary distances increase and alignment becomes more difficult (Fig. 3, A to E, close-intermediate-distant). Although the correctness of individual insertions and deletions created decreases in more difficult alignments (table S1), the phylogeny-aware method still suffers no systematic bias concerning the number of each type of event inferred.

As errors increase with greater evolutionary distances, the only way to improve alignments would seem to be to follow the practice widely used in phylogenetics, that is, to sample additional intermediate sequences (*17*–*19*), which increases the average sequence similarity. We find that this additional sequence information does not help the traditional methods.

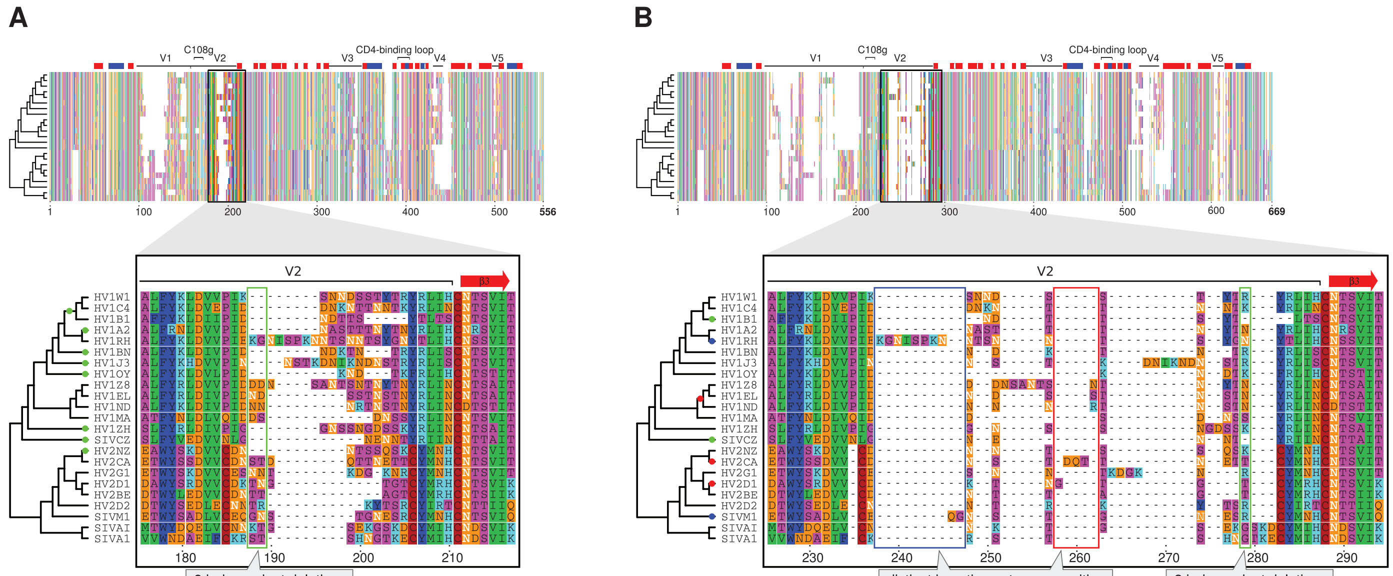

Fig. 1. Different sequence alignment approaches can give contradicting pictures of evolutionary mechanisms behind functional sequence changes. (**A**) (top) The CLUSTAL W (*12*) alignment of gp120 from different strains of human and simian immunodeficiency virus (*5*) represents a typical structural matching of protein sequences and clusters nearby alignment gaps in tight blocks. (bottom) The expanded fragment suggests that part of the V2 region evolves with a high rate of point substitutions and is shortened over evolutionary time by numerous overlapping deletions. For example, the pattern of gaps highlighted with a green box requires eight independent deletions, which have occurred in the lineages marked with green dots in the tree on the left. (**B**) (top) The PRANK$_{+F}$ (*5*) alignment of the same sequences suggests a markedly different evolutionary process dominated by short insertions and deletions. This phylogeny-aware algorithm separates distinct insertions that have taken place at the same positions (bottom: examples highlighted with blue and red boxes and dots) while permitting homologous sites to be deleted multiple times when the data support this (e.g., column highlighted in green). Alignment annotations indicate the V1 to V5 variable regions, C108g epitope, CD4-binding loop, N-linked glycosylation sites (residues N in bold white type), and α helices (blue blocks) and β strands (red blocks) of the known HIV-1 gp120 structure (*5*). Both alignments used the guide phylogeny generated by CLUSTAL W (left).

Instead, the additional sampling creates increasingly serious errors for all computed measures of accuracy (Fig. 3, A to F, close-2X-4X), very similar to the patterns observed with increasing evolutionary distance. This disappointing result is explained by the fact that both greater evolutionary distances and greater numbers of closely related sequences increase the total tree length, i.e., the evolutionary time spanned by the sequence sample. This increases the chances of sequences having insertions or deletions occurring at nearby positions. Algorithms ignoring phylogeny match these nearby insertions, and the use of site-specific, lowered gap penalties encourages nearby deletions to overlap even when sequence similarity may suggest the contrary, which creates gap patterns that are phylogenetically unreasonable. This error is avoided, however, by using phylogenetic information to distinguish insertions from deletions and by treating each mutation type appropriately. In contrast to all other methods tested, PRANK$_{+F}$ is able to use the additional data from denser sequence sampling to improve the accuracy of all estimates of parameters describing the insertion-deletion processes (Fig. 3, A to D, close-2X-4X).

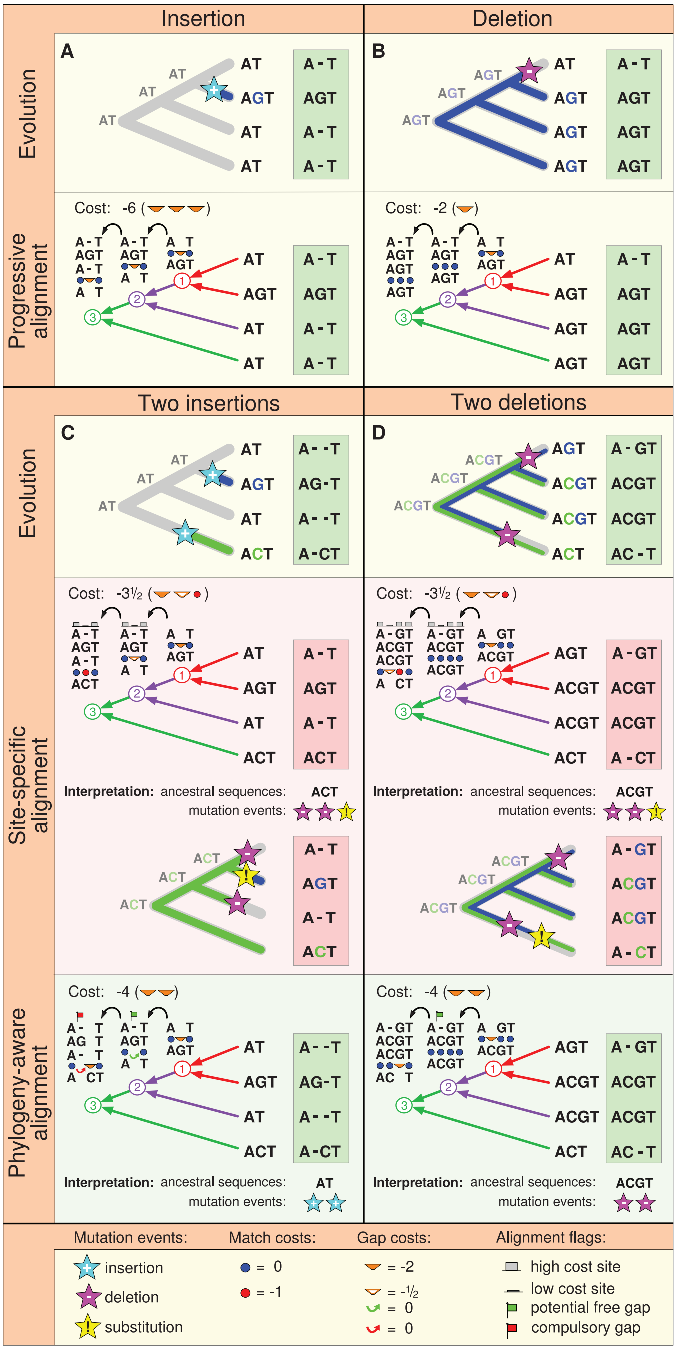

Fig. 2. Insertions and deletions are different in progressive sequence alignment. (**A**) Progressive algorithms build a multiple alignment from sequential pairwise alignments, and an insertion requires a new gap to be opened in each of them (①②③). Naïve iteration of pairwise alignment penalizes this single evolutionary event multiple times (orange triangles), giving an inappropriately high cost for the correct alignment. (**B**) A deletion is penalized only once (①, orange triangle). (**C**) A widely used "correction" for multiple penalization of insertions uses Site-specific gap penalties (gray bars) that lower the cost for reopening gaps (open triangles). With multiple nearby insertions, this encourages incorrect matching of independent events and collapses distinct insertions. Evolutionary interpretation of the incorrect alignment overestimates the substitution rate and lengths of ancestral sequences and indicates enrichment of deletions at one sequence position. The phylogeny-aware algorithm uses evolutionary information from a related sequence (②, green arrow) to confirm the earlier event as an insertion and, by marking it as permanent and preventing that site from being matched in later alignments (③, red flag), correctly places the second insertion in a column of its own and recovers the correct homology. (**D**) Site-specific lowered gap penalties create gap magnets that induce nearby deletions to coincide, also resulting in an incorrect alignment. Evolutionary interpretation again overestimates the substitution rate and indicates multiple deletions at one sequence position. The phylogeny-aware algorithm considers the earlier event (①) a deletion; removes the flag that allows for a free gap (②); and handles the second deletion correctly (③).

Wong *et al.* (*20*) showed that alignment uncertainty is crucially important in subsequent genomic analyses, such as phylogeny inference and detection of positive selection. We have shown that incorrect handling of alignment gaps is a significant contributing factor to systematic alignment error. As a further simple, but fundamental, demonstration of the effects on downstream studies, we illustrate the effect on the inference of the branch lengths of the 16-taxon intermediate phylogenetic tree. Comparing branch length estimates based on the true alignments with those based on alignments generated using the different methods, we detected patterns of errors that are consistent with our alignment accuracy results. As our analysis predicts, artifactually compact alignments with incorrect insertions and deletions create erroneous mismatches between sequence sites and cause branch lengths to be overestimated. Estimates based on the most erroneous alignments depart most significantly from the true values, and errors increase in the deeper branches (Fig. 3G). Pairwise estimates of sequence divergence and estimates of substitution rates are similarly affected. Again, branch length estimates based on the PRANK$_{+F}$ alignments were the most accurate at all levels of sequence divergence.

Our analyses show that sequence alignment remains a challenging task, and alignments generated with methods based on the traditional progressive algorithm may lead to seriously incorrect conclusions in evolutionary and comparative studies. The main reason for their systematic error is disregard of the phylogenetic implications of gap patterns created—which is not corrected by considering alignment consistency (*13*) or using post alignment refinement (*14*, *15*)—and this error is intensified by methods that intentionally force gaps into tight blocks. Affected methods can be positively misleading and become increasingly confident of erroneous solutions as more sequences are included. It is not the progressive algorithm as such that is defective, rather, correct alignment requires that we take account of sequences' phylogeny, irrespective of alignment method used or data type, but the original implementations of the progressive algorithm have a flaw that has gone unnoticed as long as different methods have been consistent in the error they create.

That such a significant error has passed undetected may be explained by the alignment field's historical focus on proteins, where these biases tend to be manifested in less-constrained regions such as loops (compare Fig. 1). Alignments with insertions and deletions squeezed compactly between conserved blocks may suffice for, and even be preferred by, some molecular biologists working with proteins. We have shown, however, that these patterns are, in fact, imposed by systematic biases in alignment algorithms, even in cases where they are incorrect and, indeed, phylogenetically unreasonable. We contend that algorithms that impose gap patterns like those found in structural alignments of proteins are inappropriate for the increasingly widespread analysis of genomic DNA and are likely to cause error when the resulting alignments are used for evolutionary inferences.

We believe that alignment methods specifically designed for evolutionary analyses will give a very different picture of the mechanisms of sequence evolution and show sequence turnover through short insertions and deletions as a more frequent and important phenomenon. This raises interesting questions of the true evolution of variable sequences such as promoter regions, noncoding DNA, and exposed coil regions in proteins: Do they predominantly evolve through point substitutions, or are those dissimilar regions just incorrectly aligned nonhomologous sequences? To resolve that, we need more sequence data and alignment methods that can really benefit from the additional information. The resulting alignments may be fragmented by many gaps and may not be as visually beautiful as the traditional alignments, but if they represent correct homology, we have to get used to them.

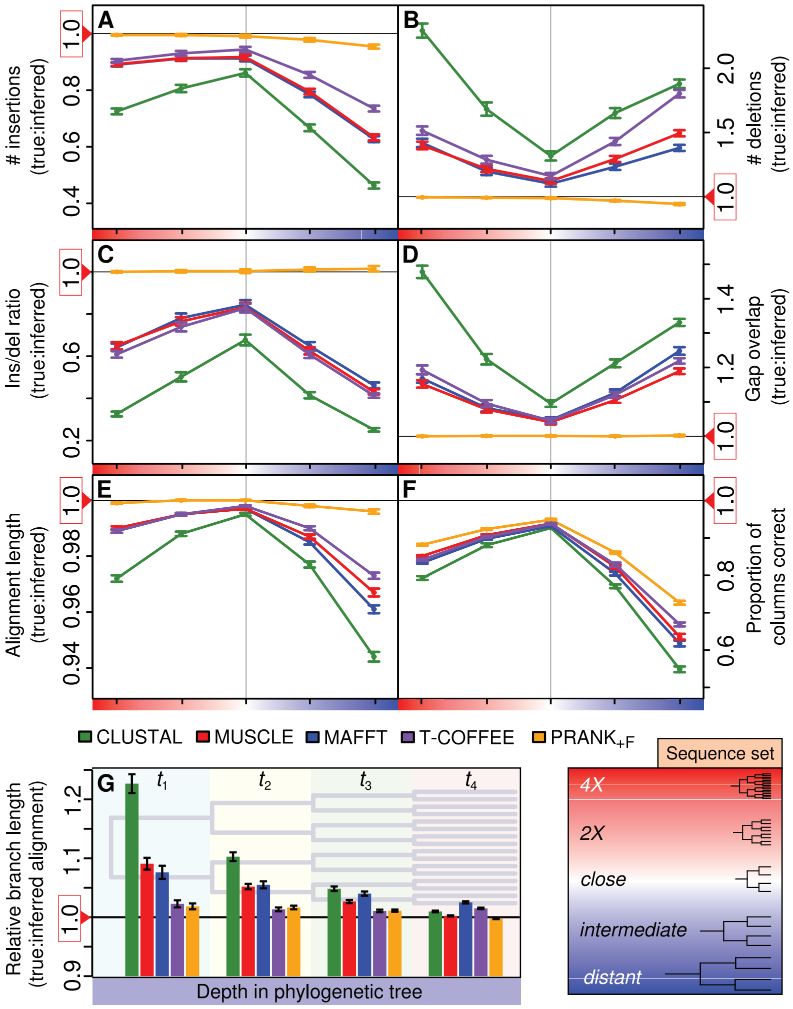

Fig. 3. Alignment accuracy errors are reduced by a phylogeny-aware algorithm. (**A** to **F**) Errors from traditional alignment methods grow with increasing evolutionary distances and more difficult alignments (close-intermediate-distant, white to blue gradient) and also with a denser sequence sampling and increasingly similar sequences (close-2X-4X, white to red gradient). The phylogeny-aware method PRANK$_{+F}$ is less biased by greater distances and, in contrast to other methods, improves in accuracy with additional sequence data. Alignment statistics for the five multiple sequence alignment methods are number of (A) insertions and (B) deletions, (C) insertion/deletion ratio, (D) gap overlap, (E) total length of the alignment, and (F) proportion of columns correctly recovered. (**G**) Inferred branch lengths at different depths in the tree ($t_1 - t_4$) for the intermediate sets indicate that alignment errors lead to overestimated branch lengths, with PRANK$_{+F}$ giving the most accurate estimates across the whole range of depths. All measures in (A) to (G) are shown relative to those inferred from the true alignment, and values closer to 1 are more correct. Vertical bars show means and 95% confidence intervals.

References and Notes

1. R. A. Gibbs *et al.*, *Nature* **428**, 493 (2004).
2. Rhesus Macaque Genome Sequencing and Analysis Consortium, *Science* **316**, 222 (2007).
3. The ENCODE Project Consortium, *Nature* **447**, 799 (2007).
4. A. Stark *et al.*, *Nature* **450**, 219 (2007).
5. Materials and methods are available as supporting material on *Science* Online.
6. A. Rambaut, D. Posada, K. Crandall, E. Holmes, *Nat. Rev. Genet.* **5**, 52 (2004).
7. N. Sullivan, M. Thali, C. Furman, D. Ho, J. Sodroski, *J. Virol.* **67**, 3674 (1993).
8. R. Wyatt *et al.*, *J. Virol.* **69**, 5723 (1995).
9. M. Jansson *et al.*, *AIDS Res. Hum. Retroviruses* **17**, 1405 (2001).
10. S. D. Frost *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18514 (2005).
11. M. Sagar, X. Wu, S. Lee, J. Overbaugh, *J. Virol.* **80**, 9586 (2006).
12. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
13. C. Notredame, D. G. Higgins, J. Heringa, *J. Mol. Biol.* **302**, 205 (2000).
14. R. C. Edgar, *BMC Bioinformat.* **5**, 113 (2004).
15. K. Katoh, K. Kuma, H. Toh, T. Miyata, *Nucleic Acids Res.* **33**, 511 (2005).
16. A. Löytynoja, N. Goldman, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10557 (2005).
17. D. D. Pollock, D. J. Zwickl, J. A. McGuire, D. M. Hillis, *Syst. Biol.* **51**, 664 (2002).
18. M. S. Rosenberg, S. Kumar, *Syst. Biol.* **52**, 119 (2003).
19. M. S. Rosenberg, *BMC Bioinformat.* **6**, 278 (2005).
20. K. M. Wong, M. A. Suchard, J. P. Huelsenbeck, *Science* **319**, 473 (2008).
21. This work was funded in part by a Wellcome Trust Programme Grant (GR078968). We thank N. Luscombe for many suggestions that improved the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/320/5883/1632/DC1
Materials and Methods
SOM Text
Fig. S1
Table S1
References

28 March 2008; accepted 15 May 2008
10.1126/science.1158395