BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 10 2012, pages 1336–1344
doi:10.1093/bioinformatics/bts158

Sequence analysis Advance Access publication April 6, 2012

PhyLAT: a phylogenetic local alignment tool

Hongtao Sun* and Jeremy D. Buhler

Department of Computer Science and Engineering, Washington University, Saint Louis, MO, 63130, USA

Associate Editor: David Posada

ABSTRACT

Motivation: The expansion of DNA sequencing capacity has enabled the sequencing of whole genomes from a number of related species. These genomes can be combined in a multiple alignment that provides useful information about the evolutionary history at each genomic locus. One area in which evolutionary information can productively be exploited is in aligning a new sequence to a database of existing, aligned genomes. However, existing high-throughput alignment tools are not designed to work effectively with multiple genome alignments.

Results: We introduce PhyLAT, the phylogenetic local alignment tool, to compute local alignments of a query sequence against a fixed multiple-genome alignment of closely related species. PhyLAT uses a known phylogenetic tree on the species in the multiple alignment to improve the quality of its computed alignments while also estimating the placement of the query on this tree. It combines a probabilistic approach to alignment with seeding and expansion heuristics to accelerate discovery of significant alignments. We provide evidence, using alignments of human chromosome 22 against a five-species alignment from the UCSC Genome Browser database, that PhyLAT's alignments are more accurate than those of other commonly used programs, including BLAST, POY, MAFFT, MUSCLE and CLUSTAL. PhyLAT also identifies more alignments in coding DNA than does pairwise alignment alone. Finally, our tool determines the evolutionary relationship of query sequences to the database more accurately than do POY, RAxML, EPA or pplacer.

Availability: www.cse.wustl.edu/~htsun/phylat

Contact: sunhongtao@wustl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 9, 2011; revised on March 29, 2012; accepted on March 30, 2012

1 INTRODUCTION

The proliferation of high-throughput DNA sequencing has produced a huge amount of genomic DNA sequence, including genomes from many higher eukaryotes. Intensively studied clades of organisms—such as mammals, *Drosophila* fruit flies and worms of the genus—are now typically represented in public databases by complete or partial genomes of multiple species. A group of closely related genomes can be combined into a large-scale multiple alignment of orthologous sequences (Blanchette, 2007; Blanchette *et al*., 2004; Wheeler and Kececioglu, 2007).

Building a high-quality multiple-genome alignment requires a large investment of computational resources and curation time, particularly if the alignment will become a reference for future users. We would therefore like to amortize this investment by effectively utilizing information present in the alignment that is not readily available from its component genomes. Multiple-genome alignments are commonly used to interrogate a clade's evolutionary history (Cliften *et al*., 2003), often with the help of a phylogenetic tree on the component species, or to discover genomic loci of unusually high conservation (Bejerano *et al*., 2004) or unusually fast change (Bird *et al*., 2007). However, they are rarely used to augment one of the most common operations in bioinformatics: aligning a new sequence to an existing reference.

In principle, using a reference multiple alignment, rather than any one of its component genomes, to align a query sequence should result in a more accurate alignment, since the aligner can use the pattern of conservation at each position to more accurately determine which query base corresponds to which multiple alignment column. Moreover, given a phylogenetic tree relating the species in the reference, an aligner should be able to use standard probabilistic models of evolution to compare the likelihoods of possible alignments, rather than resorting to an arbitrary scoring system. In fact, alignment could even infer the evolutionary relationship of the query to the reference, placing it on the tree of the reference's species.

In practice, however, most widely used alignment tools either cannot use reference multiple alignments or cannot do so in a phylogenetically aware way. BLAST (Altschul *et al*., 1997) and other accelerated variants of Smith–Waterman (Smith and Waterman, 1981) are widely used for pairwise sequence comparison, but these methods compare only individual sequences. PSI-BLAST (Altschul *et al*., 1997) creates an alignment between a query and a profile computed from a database of individual reference sequences. However, the construction of the profile does not take into account phylogenetic information, so it does not weigh each reference sequence in a phylogeny-aware way. HMMER (Eddy, 1995, 1998; Eddy *et al*., 1995) and SAM (Karplus *et al*., 1998) *can* align a sequence to a pre-existing multiple alignment, if it is generalized to a profile hidden Markov model, but even these tools use only statistical conservation at each position, rather than phylogenetic information, to perform alignments.

Tools for *de novo* multiple sequence alignment exist that use trees to improve alignment quality. The classic CLUSTAL (Thompson *et al*., 1994) software uses a guide tree to align multiple sequences. More recent tools, such as POY (Varón *et al*., 2010), can align sequences given a tree or jointly compute a multiple alignment and a supporting phylogeny. However, these tools cannot as a rule incrementally update a multiple alignment and tree starting from

*To whom correspondence should be addressed.

pre-existing references, which is the computation needed to align a query sequence to a multiple alignment database. Moreover, the high cost of *de novo* multiple alignment limits the computationally feasible methods that these tools can employ. PaPaRa (Berger and Stamatakis, 2011), unlike the tools described above, uses a guide tree to map queries, in particular short reads, to an existing multiple alignment, but it does not score or improve the resulting alignments probabilistically given the phylogeny. Practical implementation of high-throughput pairwise alignment between a query sequence and database of reference multiple alignments with phylogenetic information therefore remains an open problem.

Classical results from phylogeny (Yang, 1995) give the theory needed to construct a maximum-likelihood alignment between a query sequence and a reference multiple alignment, provided that neither query nor reference contains gaps. This theory can be extended to allow the query to contain bases that are not homologous to any reference position or vice versa; for example, Siepel and Haussler describe such an approach for phylogenetic HMMs (Siepel and Haussler, 2004). However, multi-genome reference alignments typically include columns with *both* bases and gaps, which may in fact be homologous to certain query bases. Finding a reasonable way to evaluate the likelihood of such putative homologies is a difficult problem. This fundamental issue, as well as assorted technical details needed to adapt any alignment algorithm to BLAST-like high-throughput use, make the construction of a fast, phylogenetically informed tool non-trivial.

In this work, we describe PhyLAT (the Phylogenetic Local Alignment Tool), a tool for rapidly aligning a query DNA sequence to a database of multi-genome reference alignments. PhyLAT combines BLAST-style seeding and extension heuristics with a EM-like, phylogenetically aware back-end alignment algorithm. We score alignments to references containing gaps using a model that is simplified enough for efficient implementation but disallows alignment hypotheses that are demonstrably impossible alignment given the pattern of gaps in the reference. We show that PhyLAT produces results in protein-coding regions of mammalian genomes that are better supported by external evidence than the results of pairwise alignment, and that our tool can accurately infer the evolutionary relationship of the query to the species in the multiple alignment.

2 RESULTS

PhyLAT is built around an EM-like algorithm that simultaneously computes an alignment between a query sequence and a multiple alignment and predicts the placement of the query on the tree associated with the multiple alignment. The algorithm iteratively refines query alignment and branch placement until both have converged. To accelerate this core alignment algorithm, we adopt a BLAST-like seed generation and extension heuristic (see Supplementary Materials). We use the evolutionary consensus sequence of the multiple alignment to rapidly generate pairwise seed alignments, filter these seeds by E-value, and finally apply the core algorithm to each seed. The structure of the aligner is illustrated in Figure 1.

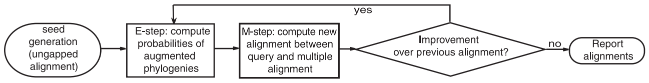
seed generation (ungapped alignment)
E-step: compute probabilities of augmented phylogenies
M-step: compute new alignment between query and multiple alignment
Improvement over previous alignment?
yes
no
Report alignments

Fig. 1. Structure of PhyLAT algorithm. Not shown is the offline preprocessing of the database to compute its consensus and parameterize a mutation model at each position.

2.1 Problem formulation for final alignment stage

Let M be a database composed of a multiple alignment of n orthologous DNA sequences. The species from which the DNA sequences are drawn are related by a phylogenetic tree τ, whose n leaves correspond to the n species. Each branch i in the tree has a length l_i, which is the evolutionary distance between the two endpoints of the branch. To convert these branches to transition probabilities, we use a mutation rate matrix Q, similar to the extended Tamura–Nei model (McGuire *et al.*, 2001; Tamura and Nei, 1993), that we estimate from the columns of M. We chose this Tamura–Nei-like model because it has a simple form with few parameters to estimate. In fact, PhyLAT can use arbitrary, non-time-reversible mutation models.

Given a query DNA sequence q, we want to find all high-scoring local alignments between q and M. We use a seed-and-extend procedure, described in the next section, to choose short substrings of q and subregions of M to align. For each such chosen pair, an alignment A is chosen to maximize a likelihood

$$\Pr(q, M|A, \tau).$$

Here, we assume that all possible alignments of q and M are a priori equally likely and choose the most likely one given the data and the tree.

Computing the complete-data likelihood for an alignment A requires that we know where the query is placed on τ relative to the sequences of M. We assume that we do not have this information; instead, we sum over all possible augmented tree topologies τ_i^* that add the query to a given branch on τ, as shown in Figure 2:

$$\Pr(q, M|A, \tau) = \sum_i \Pr(q, M, \tau_i^*|A, \tau).$$

For compactness of notation, we drop the explicit dependence of $\Pr(q, M|A, \tau)$ on the fixed tree τ in subsequent sections.

2.2 EM computation of optimal local alignment

Alignment of a query to a reference starts with *seed generation*, which produces initial ungapped *seed alignments* between the query and one or more reference regions. Details of seed alignment generation are given in Supplementary Materials. For each seed alignment, we perform gapped extension, described below. Initially, we apply this final alignment stage to a region of the query and database of length 20. To allow for final alignments of varying lengths, we retry the computation on regions whose size

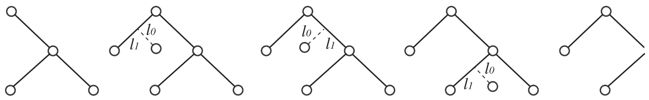

Fig. 2. An example of augmented phylogenies. The phylogeny on the left is the original, whereas the rest are its four possible augmented phylogenies. Each augmented phylogeny is actually a family of trees with two parameters l_0 and l_1, which are the length of the new branch and its attachment site to its parent branch.

is progressively doubled until doing so does not improve the final alignment score.

We now describe the EM algorithm used to compute the final local alignment A for given regions of q and M, as well as the probabilities of the augmented phylogenies τ_i^*. First, we define a set of indicator variables $\{x_i\}$ for each possible augmented phylogeny:

$$x_i = \begin{cases} 1 & \text{if augmented topology is } \tau_i^* \\ 0 & \text{otherwise.} \end{cases}$$

In this EM model, the known data are M and q (and the tree τ), while the latent variables are the x_is. The unknown parameter of the model is the alignment A. The EM algorithm iteratively refines an initial guess $A^{(0)}$ at the alignment A while simultaneously inferring a distribution over the position of q in the phylogeny. The m-th iteration starts with an alignment $A^{(m-1)}$ computed in the previous iteration. In the E-step of the iteration, the algorithm computes the expectation of each x_i:

$$\hat{x}_i = \Pr(x_i = 1 | q, M, A^{(m-1)}). \tag{1}$$

In the M-step, the algorithm computes a new alignment $A^{(m)}$ to maximize the expected log-likelihood function:

$$A^{(m)} = \underset{A}{\arg\max} \sum_i \hat{x}_i \log \Pr(q, M | x_i = 1, A). \tag{2}$$

Each iteration improves the likelihood of A and recomputes the distribution of the query's position in the tree. Finally, a local optimal point is reached, and the algorithm reports both a final alignment and an associated probability distribution over possible augmented tree topologies.

Assuming that the residues of q are stochastically independent, as are the columns of M, we can decompose the probability of the data given a tree placement and alignment as

$$\Pr(q, M | x_i = 1, A) = \prod_{j=1}^{|A|} \Pr(y[j], Z[j] | x_i = 1) \Pr_{y \notin q'}(y) \Pr_{Z \notin M'}(Z | \tau), \tag{3}$$

where $y[j]$ and $Z[j]$ are a residue in q and column in M, respectively, from the j-th column of alignment A, and q' and M' are the aligned regions of q and M, respectively.

2.3 Computation of per-column probability

In both the E-step and the M-step, we need to compute the probability of an aligned query position and multiple alignment column given an augmented tree. The details of how this probability is computed determine the accuracy and efficiency of our algorithm. We introduce two key innovations for this task: treatment of alignment gaps in a way that is informed by the tree τ, and caching of subtree probabilities to accelerate the computation.

Further details of the per-column computation beyond the highlights in this section are given in Section 3 and in Supplementary Materials.

2.3.1 Treatment of gaps An alignment of a query q to multiple alignment M may include gaps in either of q or M, or it may align a base of q to a column of M that contains both bases and gaps. To efficiently estimate the probabilities associated with such alignments, we need a gap model that is fast yet incorporates meaningful information about the alignment M. We consider two kinds of gap. The first kind, the 'local' gap, is assumed to arise as a series of single-base indels, whereas the second, the 'global' gap, arises through a mutation that adds, deletes or moves many contiguous bases at once. Local gaps are modeled using a single-base indel model, whereas global events may require a more complex model. In this work, we use the local model for sequence gaps of length ≤ 20; gaps longer than this are treated as missing data in the species where they occur. We note that this threshold was not empirically tuned to our test data but rather was an a priori estimate of the threshold between local and global indel events.

A very simple local indel model used in some work, including our own earlier work on PhyLAT (Buhler and Nordgren, 2005), treats a gap as a fifth residue that can freely interconvert with A, C, G and T. However, such treatment is inappropriate because, when we score an alignment column using a phylogeny, all observed residues are at the leaves in the tree. To compute the probability of the column, we sum over all possible labelings of the tree's internal nodes, which describe possible histories of insertion and deletion. Unfortunately, these labelings may include some histories that are biologically meaningless because they imply that aligned residues are non-homologous. The models of (McGuire *et al.*, 2001) also have the problem of illegal labelings.

Other models exist that consider only legal indel histories for a given phylogeny (Chindelevitch *et al.*, 2006; Diallo *et al.*, 2006, 2007). However, the tools using these models are computationally expensive in practice because they enumerate all possible labelings. Moreover, these models consider only whether there is a base or gap at an inner node, disregarding the identity of the base. The model of Thorne *et al.* (1991) considers only legal labelings, but it still requires a time-reversible mutation model.

The current version of PhyLAT uses a gap model that recognizes that gaps cannot interconvert freely with residues in a phylogeny. Our model imposes two constraints. (i) Once a residue is deleted (converted to gap) on a branch, it cannot later be inserted, because the inserted residues are not homologous to the original residue (see Fig. 3A). (ii) If any internal node of the tree has a gap, then only one of its children can have a residue (insertion); the other one must have a gap (see Fig. 3B). Note that once a residue is inserted, it can afterwards be deleted.

PhyLAT's per-column probability computation, while based on a simple mutation rate matrix Q (described in Section 3) that nominally treats a gap as a fifth residue, sums over *only* those configurations of internal residues that are consistent with the two constraints given above. This excludes impossible indel histories that would otherwise contribute to the computed alignment probability.

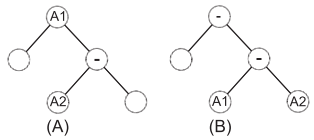

Fig. 3. (**A**) If residue $A1$ is deleted and $A2$ is then inserted, $A1$ and $A2$ should not be considered homologous. This case is not allowed in our model. (**B**) If insertions occur on both child nodes, then residues $A1$ and $A2$ should not be considered homologous. This case is not allowed in our model.

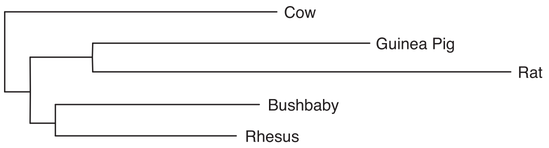

Fig. 4. Phylogeny of the species in the multiple alignment. Branch lengths are proportional to evolutionary distances.

Details of the sum computation are given in Supplementary Materials.

2.4 Experimental results

We have implemented PhyLAT in the C++ language, using the TAO optimization package (Benson *et al.*, 2007) to estimate maximum-likelihood values for the edge-length parameters l_0 and l_1 shown in Figure 2. In this section, we interrogate the result quality of PhyLAT.

We tested PhyLAT's accuracy on three queries: human chromosome 22, *Caenorhabditis elegans* chromosome 3, and *Drosophila melanogaster* chromosome 4, each aligned to a database of multiple genome alignments for related species. Here, we present only the human results, which are representative of our tool's qualitative performance versus competing aligners; the other experiments' results are described in Supplementary Material. We aligned human chromosome 22 (assembly hg19, GRCh37) against a whole-genome alignment of five mammals (shown in Fig. 4) from the UCSC genome database (Blanchette *et al.*, 2004; Kent *et al.*, 2003), which was assembled using human chromosome 22 as the reference sequence. We also tested the accuracy of tree placement by aligning opossum to a different five-species tree from the UCSC database.

2.4.1 Accuracy of DNA alignment of human chromosome 22 Currently, there are no good methods to evaluate absolute accuracy of arbitrary multiple alignments of DNA sequence (Kim and Sinha, 2007; Kumar and Filipski, 2007; Prakash and Tompa, 2007). However, protein-coding regions are generally stable and can be translated to protein and aligned by a protein aligner, producing alignments of generally higher quality than those obtained from DNA alone. We therefore validated PhyLAT's alignment quality by examining local alignments involving annotated coding sequences. We used the UCSC database's reference alignment between human and the other five species as our ground truth for orthology relationships among our sequences.

Table 1 illustrates one benefit of using multiple species for recovering alignments between the query and *orthologous* sequences in the database. We divided the alignments found by PhyLAT according to the number of species (up to five) with sequence at the locus of the alignment in the database. The more species present in the database at a given locus, the higher the probability that an alignment at that locus aligns the query to orthologous sequences. We note that alignments with more species present are not systematically better-conserved than those with fewer species; indeed, aligned regions with only two aligned sequences had higher identity on average than those with three or four. Nevertheless, the fraction of query sequences aligned to their orthologous regions in the multiple alignment increased monotonically with the number of species present.

Table 1. Effect of using multiple alignment on improving orthology detection

No. of sequences	No. of alignments	No. of orthologous	Orthologous (%)	Identical (%)
2	982	797	81.16	53.65
3	383	316	82.51	46.89
4	553	477	89.49	52.89
5	726	704	96.97	59.38
Total	2644	2294	86.76	53.83

No. of sequences, number of species present in the aligned multiple alignments; *no. of alignments*, number of PhyLAT alignments; *no. of orthologous*, number of orthologous PhyLAT alignments; *orthologous (%)*, no. of orthologous/no. of alignments; *identical (%)*, percentage of PhyLAT columns containing identical bases. The more database species present at a locus, the greater the percentage of alignments involving orthologous sequences.

2.4.2 Validation of DNA alignment by protein alignment To estimate the likely accuracy of PhyLAT's alignments in coding DNA, we extracted and translated the sequences it aligned, then used a protein multiple aligner to realign them, and finally checked whether the DNA alignment inferred from the aligned proteins matched PhyLAT's alignment. Because protein alignment uses information not available to a DNA aligner, we expect that it will yield more accurate results in general; hence, concordance between the DNA and protein alignments acts as a proxy for the (unknown) absolute accuracy of the DNA alignment.

From PhyLAT's DNA alignments involving orthologous sequences, we first extracted those portions that covered protein-coding regions (as annotated in the UCSC database). For each such alignment between a DNA query q and a multiple alignment of k DNA sequences $s_1 \dots s_k$, PhyLAT's output induces pairwise protein alignments A_i between the translation of q and that of each s_i. We compared the induced alignments A_i to alignments A'_i obtained by first translating q and s_i independently, then aligning the two resulting protein sequences using a protein-specific alignment tool. A codon in a query was considered 'accurately aligned' to the database if and only if, for $1 \le i \le k$, A_i agreed with the corresponding, independently derived A'_i over that codon. We repeated this experiment using four different protein aligners—ClustalW (Thompson *et al.*, 1994), DIALIGN (Morgenstern *et al.*, 1998), Muscle (Edgar, 2004) and T-Coffee (Notredame *et al.*, 2000)—and obtained substantially similar results with each. Additional validation would be possible by comparing our results to, e.g. structural superposition of the aligned proteins. However, such superpositions are already known to agree closely with protein aligners' output on mammalian proteins (Berman *et al.*, 2000), so we did not pursue this extra validation step.

The first part of Table 2 shows PhyLAT's accuracy on our test set using ClustalW as the protein aligner. Over 97% of query codons aligned by the algorithm were accurately aligned to the database by our measure. Moreover, PhyLAT's alignments covered >99% of all annotated codons in the multiple alignment, so this accuracy applies to essentially all the coding sequence that could possibly be aligned.

We further subdivided the protein-coding region of the database to identify regions where the protein multiple alignment induced by the DNA multiple alignment of $s_1 \dots s_k$ was inconsistent with the result obtained by independently translating each of $s_1 \dots s_k$, then aligning the resulting sequences using a protein multiple aligner. Such regions are more likely to be misaligned in the database, which in turn provides bad information to PhyLAT's aligner. For the 84% of codon positions in the database that were consistent by the above criterion, PhyLAT's accuracy was well over 99%.

An alternative to PhyLAT's approach would be to align the query to a single, representative DNA sequence instead of a DNA multiple alignment. For example, one might align our human query to one of the multiple alignment's component species' genomes, or to the evolutionary consensus of these genomes given the tree. We therefore investigated whether such pairwise alignments, as realized by the widely used BLAST software (v2.2.23+; Camacho *et al*. 2009), could match the accuracy and coverage obtained by PhyLAT.

It was not computationally feasible to BLAST the entirety of human chromosome 22 at once against a database sequence as long as our multiple alignment. Instead, for each homologous PhyLAT alignment of query segment q and database segment M, we extracted the collinear block B in UCSC's multispecies multiple alignment that contained q and M. We then used BLAST to align q to each individual sequence in B, or to its evolutionary consensus. If BLAST returned more than one local alignment between a query and a block, then we retained all such alignments. Finally, we evaluated the collection of induced BLAST alignments in protein-coding regions of the query using the same accuracy and coverage measures described above. Note that accuracy for a pairwise BLAST alignment of two coding DNA sequences is determined by agreement with a single pairwise protein alignment between them, whereas for PhyLAT, *all* induced pairwise alignments must agree with the protein aligner's results.

The second part of Table 2 shows the results of using BLAST pairwise alignments, rather than PhyLAT's approach, on our human to mammalian alignment task. For species other than rhesus, the closest to the human query, per-codon accuracy of the pairwise alignments was inferior to PhyLAT's. Aligning to the consensus actually lowered accuracy compared with two-species alignments. Moreover, the pairwise alignment sets covered fewer codons in the original multiple alignment than did PhyLAT's output. This lower coverage arises because not all species had sequence at every point in the reference multiple alignment. Hence, even aligning human to rhesus, which produced alignments to ∼100% of the codons in the rhesus sequence, yielded <92% coverage of all codons in the multiple alignment.

Overall, PhyLAT produced alignments with accuracy comparable to using BLAST to search against the best single reference species from the multiple alignment, while offering substantially improved coverage because of the availability of multiple species to cover assembly or homology gaps left by any one species' genome.

We also compared PhyLAT with other commonly used multiple alignment tools, including POY, MAFFT (Katoh *et al*., 2002), MUSCLE, CLUSTAL and PaPaRa (Berger and Stamatakis, 2011). Because these programs produce multiple alignments, which include all input sequences, it is not proper to feed the whole reference sequences and query to them to produce genome-scale multiple alignments. Instead, we use homologous segments from the

Table 2. Comparison among PhyLAT alignments, BLAST pairwise alignments and alignments of other phylogeny-aware tools

Program	Database	No. of aligned	No. of accurate	Accuracy (%)	No. of total	Coverage in species (%)	Coverage in MA (%)
PhyLAT	Whole MA	16 404	15 956	97.27	16 445	–	99.75
	Consistent MA	13 487	13414	99.46	–	–	–
	Inconsistent MA	2917	2542	87.14	–	–	–
BLAST	Cow	12 384	11 607	93.73	15 129	81.86	75.31
	Guinea pig	12 306	11 667	94.81	16 159	76.16	74.83
	Bushbaby	10 732	9748	90.83	12 527	85.67	65.30
	Rhesus	15 066	14 700	97.57	15 077	99.93	91.61
	Rat	11 503	10 597	92.12	15 787	72.86	69.95
	Consensus	16 174	14 420	89.16	16 445	–	98.35
POY	Whole MA	16 445	14 556	88.51	16 445	–	100.00
	Consistent MA	13 502	12 423	92.01	–	–	–
	Inconsistent MA	2943	2133	72.48	–	–	–
MAFFT	Whole MA	16 445	15 600	94.86	16 445	–	100.00
	Consistent MA	13 502	13 122	97.19	–	–	–
	Inconsistent MA	2943	2442	82.98	–	–	–
MUSCLE	Whole MA	16 445	15 181	92.31	16 445	–	100.00
	Consistent MA	13 502	12 911	95.62	–	–	–
	Inconsistent MA	2943	2270	77.13	–	–	–
CLUSTAL	Whole MA	16 445	15 238	92.66	16 445	–	100.00
	Consistent MA	13 502	13 044	96.61	–	–	–
	Inconsistent MA	2943	2194	74.55	–	–	–
PaPaRa	Whole MA	16 445	15 296	93.01	16 445	–	100.00
	Consistent MA	13 502	13 091	96.96	–	–	–
	Inconsistent MA	2943	2205	74.92	–	–	–

No. of aligned, total no. of codons in query aligned to the database; *no. of accurate*, number of aligned codons in previous column that are aligned the same by DNA and protein aligners; *accuracy*, ratio of accurate to aligned codons; *no. of total*, total number of codons present in the indicated sequence; *coverage in species*, total number of codons in species' sequence covered by query alignments; *coverage in MA*, total number of codons in entire multiple alignment database covered by query alignments.

reference sequences and the query where PhyLAT finds alignments. The results are shown in Table 2. Note that because we do global alignments on the input, these aligners aligned all the input codons.

2.4.3 Tree placement of human sequences Another measure of PhyLAT's accuracy is whether it placed each query sequence in its correct location on the tree of the species in the database. For the local alignments of orthologous sequences in our test set, EM should place the human query in its accepted location relative to the other, non-human mammalian species with a high-posterior probability while assigning low probabilities to incorrect placements.

Although some methods are available for comparing two trees with branch lengths (Pattengale *et al*., 2007), there is not an acknowledged standard on correct branch lengths for a given phylogeny. We therefore assessed only whether the most likely placement of the query in each local alignment was topologically correct, i.e. was the human sequence placed on the branch leading to rhesus, or on some other branch?

Figure 5 shows how many alignments placed the human sequence on each branch of the phylogeny with highest probability. In total, 88.7% of alignments correctly placed the human sequence adjacent to rhesus. If we add in 'almost correct' placements (defined as branch placement adjacent to the correct one), the fraction of such placements rises to 95.2%.

We also compared PhyLAT's accuracy with that of tools whose results include a branch placement for the query sequence, including POY, RAxML (Stamatakis, 2006), EPA (Berger *et al*., 2011), pplacer (Matsen *et al*., 2010) and PaPaRa (Berger and Stamatakis, 2011). Because all these programs need multiple alignments to do prediction, we used only orthologous informant sequences and queries as the input. Because gene trees may be different from species trees, in order to assess branch placement accuracy, we also divided the placements into two categories: those whose informant trees matched the trees built by PHYLIP (Felsenstein, 1989) from MAFFT alignments of the sequences, and those whose informant trees were incongruent with their PHYLIP trees. The results are shown in Table 3. PhyLAT's placement accuracy was substantially greater, both absolutely and relative to its competitors, when the informant sequences matched the supplied phylogeny. We note that POY has only one correct placement; this is because it builds an entirely new tree on the input sequences instead of just inserting the query species into the existing tree of informant species. We provided the informant phylogeny as input restrictions on the tree topology, but POY used it only as a starting tree and failed to produce output trees consistent with these restrictions.

We further investigated how confident PhyLAT typically was about its branch placements. A confident placement has the vast majority of the probability mass, with little probability assigned to other hypotheses, while a low-confidence placement distributes the probability more equally across branches. We computed the entropy for the posterior placement distribution of each query, summarizing these entropies in a histogram in Figure 6. For most correct branch placements, PhyLAT was highly confident about its predictions, while confidence for incorrect predictions was typically lower. For almost-correct placements, the ratio of high- to low-confidence placements is close to even. We could detect and reject most incorrect placements, with relatively few false rejections, by rejecting any placement with an entropy over 0.25 bits.

Table 3. Tree placement of orthologous human query sequences

Program	No. of correct in congruent	No. of correct in incongruent	No. of total correct	overall accuracy (%)
PhyLAT	1452	641	2093	91.72
POY	1	0	1	0.04
RAxML	669	599	1268	55.57
EPA	650	634	1284	56.27
pplacer	731	695	1426	62.49

There are 1558 congruent informant trees and 641 incongruent informant trees.

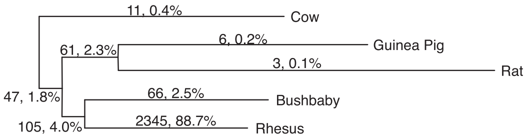

Fig. 5. Tree placements for all human query sequences. The correct location of the query is on the branch leading to rhesus. Each branch is labeled with the number of queries placed on that branch, as well as the percentage of all queries that this number represents.

The absolute accuracy of tree placement for our experiments on *C. elegans* and *D. melanogaster* was considerably lower than for our mammalian alignment—between 40% and 50%. However, as in the mammalian case, PhyLAT's results were more accurate than those of competing tools that gave placement information in their output. Details may be found in Supplementary Material.

2.4.4 Tree placement of the opossum species In spite of the fact that phylogenetic relations of existing species have been explored extensively, many relations remain missing, and many are being modified constantly. In a recent update of the phylogenetic tree of 46 species from the UCSC database, 35 species changed their tree placements (Rhead, 2010). One example of such a change was the movement of opossum from relatively near the root of the mammalian phylogeny to a location much closer to other marsupials such as wallaby.

As a further test of PhyLAT on a different dataset, we aligned opossum chromosome X to a five-species multiple alignment from the UCSC genome database. PhyLAT produced 931 local alignments, with branch placements of the opossum queries as shown in Figure 7. Assuming, as in the revised UCSC tree, that the

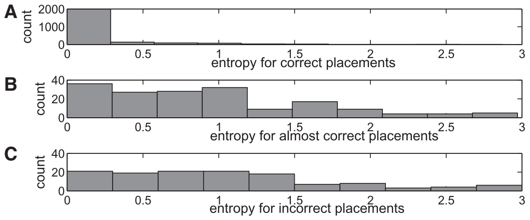

Fig. 6. Histograms of entropies (in bits) of posterior branch placement distributions. Because there are eight possible values for the branch placement, the entropy is in interval [0,3]. The smaller the entropy, the more concentrated the probability distribution, and the more confidence PhyLAT has in the branch placement.

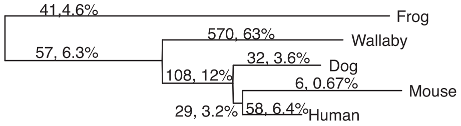

Fig. 7. Tree placement for opossum. Branch lengths are proportional to evolutionary distances. The correct location of the query is on the branch leading to wallaby. Opossum had until recently been placed on the branch leading to the parent of wallaby. Each branch is labeled with the number of queries and the fraction of all queries placed on it.

correct placement is on the branch to wallaby, 63.3% of queries were placed correctly, whereas 81.3% were placed correctly or almost correctly. In contrast, the number of queries placed at opossum's old location in the tree was only 6.3%. This example shows that PhyLAT's placement probabilities can be useful for discovering inconsistencies with an accepted phylogeny.

2.5 Computational efficiency

Our implementation of PhyLAT used several techniques to reduce the time cost. We used a compact, column-oriented storage format for the database to reduce cache misses. We used a customized phylogenetic caching techniques to store per-column probabilities, which greatly reduced the cost of probability computations. Finally, we tweaked the parameters of our algorithm for greater efficiency while maintaining high accuracy.

PhyLAT took 16.01 h to compute all alignments of human to the five-species mammalian multiple alignment on an AMD Opteron 2.4 GHz processor. In contrast, using BLAST to search human against the individual species—rhesus, rat, cow, bushbaby and guinea pig—took 3.51, 1.10, 2.11, 1.29 and 2.06 h, respectively. Using BLAST to search human against the evolutionary consensus took 3.90 h. (In all cases except the PhyLAT run, we had to break the human sequence into small pieces to avoid quadratic behavior by NCBI BLAST.) Hence, while there is a non-trivial cost to using PhyLAT, it is well within an order of magnitude of simpler pairwise alignment, while yielding superior results.

The time complexity of PhyLAT lies mainly in the E-step, i.e. optimizing branch placements. With increasing numbers of informant species, this cost comprises the greatest part in the running time. We tested with 2–10 informant species. Figure 8 shows the comparison of running time per seed.

2.5.1 Accuracy impact of additional informants We also investigated the effect of increasing informant species from 2 to 10 on alignment accuracy and branch placement accuracy. We found that, while adding distant species can increase coverage of the query sequences, it may not help increase branch placement or alignment accuracies. Because distant species are hard to align to their orthologous species during the construction of a multiple alignment, they may add noise to the search. Moreover, too-distant species will not contribute more information about where to place the query when closer species are already available. For detailed results, please see the Supplementary Materials.

2.6 Evaluation of statistical significance

It has been shown that for many alignment problems, the optimal scores under the null hypothesis empirically follow a Gumbel distribution (Altschul and Gish, 1996; Altschul *et al*., 1997, 2001; Bundschuh, 2002; Olsen *et al*., 1999; Poleksic, 2009; Prakash and Tompa, 2005; Sadreyev and Grishin, 2003). We tested this hypothesis for the scores of alignments generated by PhyLAT by running it on real queries and permuted multiple alignments. We found that the score distribution closely matched the Gumbel distribution; details are given in Supplementary Material.

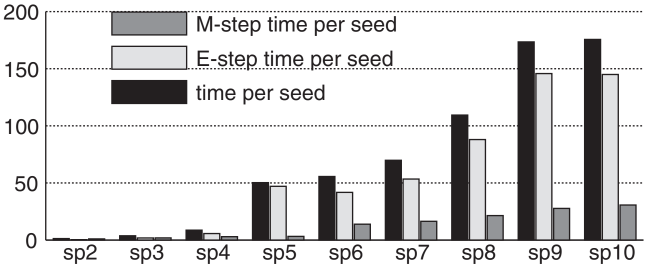

Fig. 8. Comparison of running time per seed. The Y-axis shows the running time in seconds.

3 METHODS

3.1 Parametrization of PhyLAT

PhyLAT's evolutionary model was parametrized to match the contents of the multiple alignment database, using maximum likelihood estimation. Mutation rates in the matrix Q were scaled so as to yield a total rate of one accepted mutation per unit of time.

To accommodate non-uniform evolutionary rates at different sites in the multiple alignment (Felsenstein and Churchill, 1996), we considered fast, medium and slow models at each column of the multiple alignment, corresponding to branch lengths of 0.1, 1 or 10 times the lengths given in the UCSC-supplied tree. Each column was assigned its most likely model from among these three.

The banded affine alignment algorithm used in the M-step of gapped extension used a bandwidth of 101 and an arbitrary gap-opening penalty of -12. We note that addition of a gap-opening penalty makes our aligner not strictly probabilistic, and hence not truly EM; however, we felt that this concession was worth the practical benefits of an affine algorithm, given that our probabilistic model alone would score gaps as linear rather than affine. We did *not* add an arbitrary gap-extension penalty but rather used the log-probabilities inferred for each individual gap.

3.2 Computation of E-step

To compute $\hat{x}_i$, we apply Bayes' theorem to Definition 1:

$$\hat{x}_i = \frac{\Pr(q,M|x_i=1,A^{(m-1)})\Pr(x_i=1|A^{(m-1)})}{\Pr(q,M|A^{(m-1)})}. \tag{4}$$

The first term of the numerator is given by Equation (3), whereas the second term is independent of the actual data and can be viewed as the user's *prior* over the position of q in the tree. The denominator is independent of x_i and can be normalized away.

3.3 Computation of M-step

By subtracting a constant from Equation (2), the alignment that maximizes Equation (2) is the one that maximizes

$$
\begin{aligned}
&\sum_i \hat{x}_i \log \Pr(q,M|x_i=1,A^{(m)}) - \sum_i \hat{x}_i \log \Pr(q)\Pr(M|\tau) \\
&= \sum_j \sum_i \hat{x}_i \log \frac{\Pr(y[j],Z[j]|\tau_i^*)}{\Pr(y[j])Pr(Z[j]|\tau)} \\
&= \sum_j \sigma(y,Z),
\end{aligned}
$$

where $\sigma(y,Z)$ is the per-column probability, which can be precomputed and cached.

4 DISCUSSION

While PhyLAT generally aligned query sequences to their orthologous regions as inferred from the UCSC database, there were 350 human query sequences which were not aligned to their orthologous regions. We examined these query sequences using BLAST and found that all of them are highly conserved coding regions, which are likely difficult to distinguish from the orthologous loci on the basis of similarity alone. These 350 sequences are spread over 56 genes; the most frequently hit gene was IGL2, hit by 87 queries, which is an immunoglobulin lambda locus.

We also examined 48 human queries whose branch placements in PhyLAT were not on the branch leading to rhesus yet had entropies < 0.25, indicating high confidence in the 'wrong' placement. These queries hit CDS regions from 19 genes, listed in Table 4. We searched for these genes and their families in TreeFam (Li *et al.*, 2006) and found that, for 45 of the 48 placements, the gene trees were different from the species tree. We hypothesized that PhyLAT's 'wrong' branch placement prediction in these cases was an artifact of the incorrect tree provided for the informant species. To test this hypothesis, We reran PhyLAT on the 45 identified cases using the appropriate gene tree from TreeFam as the reference tree in each case. We found that, in 40 out of the 45 cases, PhyLAT's query placements with the revised input tree were consistent with the gene tree from TreeFam.

We further note that the distribution of the 48 wrong placements is in accordance with the theory of deep coalescence, which states that short wide trees may show more genes with deep coalescence than long narrow trees (Maddison, 1997). See Figure 9 for an example.

5 CONCLUSION

We have described PhyLAT, an efficient, phylogenetically aware tool computing and scoring local alignments between a query sequence and a large multiple alignment. In our tests, PhyLAT's accuracy in alignments involving coding DNA exceeded that of BLAST, even using the consensus of all species in the database. PhyLAT's accuracy also exceeded that of phylogeny-aware alignment tools. Moreover, the inferred placement of queries on the tree of the database species was typically both topologically accurate and confident. We believe that PhyLAT's methods will prove useful in augmenting existing high-throughput sequence comparison tools to exploit the extra information provided by multiple-genome alignment databases.

Several opportunities exist to improve PhyLAT's performance and utility. First, our assumption that successive residues in the query or successive columns in the multiple alignment are stochastically independent is not realistic. It could be useful to add a dependence model between adjacent bases/columns to improve the accuracy of alignment.

Table 4. Tree placements of queries with entropy <0.25 bits

Count	Tree placement	Gene	Species = gene tree?
1	Bushbaby	IGL@	No
1	Bushbaby	LOC644525	–
1	Bushbaby	OR11H1	No
1	Bushbaby	POTEH	–
1	Parent of bushbaby	KIAA1644	No
1	Parent of bushbaby	KLHL22	No
1	Parent of bushbaby	LDOC1L	No
1	Parent of bushbaby	LOC644525	–
1	Parent of bushbaby	NCRNA00207	–
1	Parent of bushbaby	SLC5A4	No
1	Rat	IGL@	No
1	Guinea pig	NCRNA00207	–
1	Parent of rat	DEPDC5	–
1	Prent of rat	RTN4R	No
1	Cow	GRAMD4	Yes
1	Cow	TTC28AS	No
2	Parent of bushbaby	CES5AP1	–
2	Parent of bushbaby	IGKV2OR22-4	–
2	Parent of rat	CABIN1	Yes
2	Neighbor of cow	IGL@	No
2	Cow	DDTL	No
3	Bushbaby	LRP5L	No
4	Parent of bushbaby	PRAMEL	No
7	Parent of bushbaby	IGL@	No
8	Parent of rat	IGL@	No

Second, PhyLAT assumes that its query is a single DNA sequence. It would be useful to handle queries that are themselves multiple alignments. However, there are unresolved computational complexity issues with this extension. In particular, we cannot simply enumerate all simultaneous branch placements of all species in the query multiple alignment with respect to the database multiple alignment. Some efficient way must be found to form a hypothesis about how the query species and database species relate within a single tree.

Third, we need to develop efficient approaches to estimate the statistical significance of gapped alignments in PhyLAT without resorting to expensive simulations. Karlin and Altschul's theory for ungapped alignments is not applicable to gapped alignments (Karlin and Altschul, 1993). However, many kinds of alignments involving gaps were empirically demonstrated to follow EVD (Altschul *et al.*, 1997; Eddy, 2008; Sadreyev and Grishin, 2003; Schaffer *et al.*, 1999). To derive empirical parameters of KA statistics, tens of thousands of alignments need to be generated and scored. This

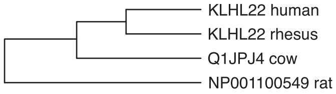

Fig. 9. Example of inconsistency between species tree and gene trees. The above shows a gene tree from TreeFam. Given the species tree shown in Figure 4, the human query is placed on the branch leading to cow. The species tree shows that the human gene is closer to rat than cow, whereas TreeFam shows that the human gene is actually closer to cow than rat.

process is computationally expensive but may not give parameters accurate enough for computing statistical significance for new alignments, especially when the composition of new data is different from that in simulation. In Sadreyev and Grishin (2003), a rescaling technique was explored to use a standard score distribution to estimate statistical significance of profile alignments of new profiles. In Poleksic (2009), the island method was applied to collect more optimal scores from a single simulated alignments. Recently it was shown that all fully-probabilistic HMM models have the property that the scores follow a Gumbel distribution (Eddy *et al*., 2009), but this is not applicable to those alignment models not based on HMM.

Finally, it would be useful to consider alternative database trees during alignment, e.g. to accommodate the possibility that a query is not being aligned to an orthologous locus in the database. While it is possible to sum probabilities over multiple tree hypotheses, the increased computational cost of using multiple trees makes it imperative to be careful not to consider too many such alternative trees. Heuristics for picking likely trees would help to guide the search.

Funding: This work was supported by National Science Foundation (NSF) award DBI-0237902 and by National Institute of Health (NIH) award R42 HG003225.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389.

Altschul,S.F. *et al*. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351.

Bejerano,G. *et al*. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321.

Benson,S. *et al*. (2007) TAO User Manual (Revision 1.9).

Berger,S.A. and Stamatakis,A. (2011) Aligning short reads to reference alignments and trees. *Bioinformatics*, **27**, 15.

Berger,S.A. *et al*. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.*, **60**, 3.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bird,C.P. *et al*. (2007) Fast-evolving noncoding sequences in the human genome. *Genome Biol.*, **8**, R118.

Blanchette,M. (2007) Computation and analysis of genomic multi-sequence alignments. *Ann. Rev. Genom. Hum. G*, **8**, 193–213.

Blanchette,M. *et al*. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708.

Buhler,J. and Nordgren,R. (2005) Toward a phylogenetically aware algorithm for fast DNA similarity search. *Lect. Notes Comput. Sci.*, **3388**, 15–29.

Bundschuh,R. (2002) Rapid significance estimation in local sequence alignment with gaps. *J. Comput. Biol.*, **9**, 243–260.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chindelevitch,L. *et al*. (2006) On the inference of parsimonious indel evolutionary scenarios. *J. Bioinform. Comput. Biol.*, **4**, 721–744.

Cliften,P. *et al*. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71.

Diallo,A.B. *et al*. (2006) Finding maximum likelihood indel scenarios. *Lect. Notes Comput. Sci.*, **4205**, 171.

Diallo,A.B. *et al*. (2007) Exact and heuristic algorithms for the indel maximum likelihood problem. *J. Comput. Biol.*, **14**, 446–461.

Eddy,S.R. (1995) Multiple alignment using hidden Markov models. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Vol. 3, pp. 114–120.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy,S.R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.*, **4**, e1000069.

Eddy,S.R. *et al*. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, **2**, 9–23.

Eddy,S.R. *et al*. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792.

Felsenstein,J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Felsenstein,J. and Churchill,G.A. (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93.

Karlin,S. and Altschul,S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci.*, **90**, 5873.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846.

Katoh,K. *et al*. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 14.

Kent,W.J. *et al*. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484.

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

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

Li,H. *et al*. (2006) TreeFam: a curated database of phylogenetic trees of animalgene families. *Nucleic Acids Res.*, **34**, D572–D580.

Maddison,W.P. (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 3.

Matsen,F. *et al*. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 1.

McGuire,G. *et al*. (2001) Models of sequence evolution for DNA sequences containing gaps. *Mol. Biol. Evol.*, **18**, 481.

Morgenstern,B. *et al*. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290.

Notredame,C. *et al*. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Olsen,R. *et al*. (1999) Rapid assessment of extremal statistics for gapped local alignment. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 211–222.

Pattengale,N.D. *et al*. (2007) Efficiently computing the robinson-foulds metric. *J. Comput. Biol.*, **14**, 724–735.

Poleksic,A. (2009) Island method for estimating the statistical significance of profile-profile alignment scores. *BMC Bioinformatics*, **10**, 112.

Prakash,A. and Tompa,M. (2005) Statistics of local multiple alignments. *Bioinformatics*, **21**, 344.

Prakash,A. and Tompa,M. (2007) Measuring the accuracy of genome-size multiple alignments. *Genome Biol.*, **8**, R124.

Rhead,B. *et al*. (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

Schaffer,A.A. *et al*. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000.

Siepel,A. and Haussler,D. (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.*, **11**, 413–428.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 21.

Tamura,K. and Nei,M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512.

Thompson,J.D. *et al*. (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.

Thorne,J.L. *et al*. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, **33**, 114–124.

Varón,A. *et al*. (2010) POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics*, **26**, 72–85.

Wheeler,T.J. and Kececioglu,J.D. (2007) Multiple alignment by aligning alignments. *Bioinformatics*, **23**, i559.

Yang,Z. (1995) A space-time process model for the evolution of DNA sequences. *Genetics*, **139**, 993.