BIOINFORMATICS LETTER TO THE EDITOR Vol. 23 no. 23 2007, pages 3258–3262
doi:10.1093/bioinformatics/btm402

Phylogenetics

Transducers: an emerging probabilistic framework for modeling indels on trees

Robert K. Bradley[1] and Ian Holmes[2,*]

[1]Department of Physics and [2]Department of Bioengineering, University of California, Berkeley CA, USA

Received on March 22, 2007; revised on August 1, 2007; accepted on August 3, 2007

Advance Access publication September 5, 2007

Associate Editor: Alex Bateman

Contact: ihh@berkeley.edu

1 INTRODUCTION

When it comes to dealing with indels, molecular evolution lags heuristic bioinformatics by decades. Sophisticated alignment algorithms have been widely known since the 1960s (and in bioinformatics since 1970), but we are still struggling to understand the corresponding phylogenetic models. Big ideas drive change: as we dream of reconstructing ancestral genotypes, it is ever clearer that indels cannot be ignored. We need to develop a robust understanding of probabilistic indel analysis and its relationship to alignment.

We believe that a suitable foundation for such analysis already exists, where evolutionary models meet automata theory: the framework of *finite-state transducers*. This framework links Hidden Markov Models (Brown *et al*., 1993; Churchill, 1992), sequence alignment algorithms (Gotoh, 1982; Miller and Myers, 1988; Needleman and Wunsch, 1970; Smith and Waterman, 1981), finite-state machines and Chomsky grammars (Durbin *et al*., 1998) and molecular phylogenetics (Miklós *et al*., 2004; Thorne *et al*., 1991). In this letter we outline this framework, also describing a preliminary analysis of one recent algorithm—*Indelign*—for reconstructing ancestral indel histories (Kim and Sinha, 2007).

Below, we briefly review the theory of transducers, concentrating not on the details of individual algorithms but rather on their unifying qualitative character. We show that Indelign, which reconstructs maximum-likelihood indel histories, is implicitly based on a transducer model. Thus, we can compare the computational complexity of Indelign to other transducer-framed algorithms, with reference to alignment data from recent comparative genomics projects in *Drosophila* and *Eutheria* (ENCODE). Finally, we discuss several programs, algorithms and resources available for working with transducers, offering an outlook on areas of bioinformatics that may benefit from this theory.

1.1 Theory of finite-state transducers

A transducer is a finite-state machine with an input tape (X), an output tape (Y), a symbol alphabet and a set of *transition* and (possibly) *emission weights*. It is therefore very similar to a Pair HMM (Hidden Markov model), which is also a two-tape finite-state machine with transition and emission weights (Durbin *et al*., 1998). As with a Pair HMM, each transducer state may be classified as Match, Insert, Delete, Start or End. In both cases, a path π through the machine corresponds to a pairwise sequence alignment with an associated likelihood, defined to be the product of transition and emission weights along the path. Pair HMMs and transducers have similar sets of algorithms for inference, including the Forward, Backward and Viterbi algorithms (Durbin *et al.,* 1998).

The crucial difference is that the Pair HMM's tapes are both considered to be outputs, whereas the transducer has one input and one output. The probabilistic interpretation is that the path probability for a Pair HMM is the joint likelihood $P(\pi, X, Y)$, while for a transducer it is the conditional likelihood $P(\pi, Y|X)$. Conceptually, *a transducer represents the operation of a finite span of evolutionary time* (ΔT), 'evolving' the input sequence into the output sequence by introducing substitutions and indels at random. We can represent this operation as $X \xrightarrow{\Delta T} Y$.

The feature of transducers that makes them so useful for comparative sequence analysis is the existence of algorithms for *composing* them in series or in parallel (Holmes, 2003; Mohri *et al.*, 2000), where a series composition represents the consecutive operation of two transducers ($X \xrightarrow{\Delta T} Y \xrightarrow{\Delta T} Z$) and a parallel composition represents a bifurcation in a phylogenetic tree ($X \xrightarrow{\Delta T} Y$ and $X \xrightarrow{\Delta T} Z$). By placing a transducer on each branch of a phylogenetic tree, we can automate the construction of systematic scoring schemes and algorithms for alignment, annotation or parameter estimation. Transducers are natural models for indels on trees, just as continuous-time Markov chains are natural models for substitutions.

Although transducers have been known in the computer science literature since the 1950s (Mealy, 1955; Mohri *et al*., 2000), they have been applied in bioinformatics only lately (Holmes, 2003; Searls and Murphy, 1955). In fact, early probabilistic alignment algorithms share similarities to transducers (Bishop and Thompson, 1986), as (of course) do Pair HMMs (Holmes and Durbin, 1998). The breakthrough came

[*]To whom correspondence should be addressed.

in the field of *statistical alignment* (a term coined by Jotun Hein), which attempts to unite bioinformatics and molecular evolution via explicit birth–death models for indels and other events. In pioneering work, the TKF91 model of Thorne *et al*. (1991) was used to derive alignment algorithms with linear gap penalties; these were then extended to multiple sequences on a tree (Hein, 2001), recognized as examples of HMM algorithms (Holmes and Bruno, 2001) and formulated using transducers (Holmes, 2003).

Although the linear gap penalty of TKF91 is occasionally quoted as a drawback of statistical alignment, this is a misconception of the role of TKF91. Several transducers with affine gap penalties have been derived from evolutionary models (Knudsen and Miyamoto, 2003; Miklós *et al*., 2004). TKF91's role can be seen as a well-studied and canonical (albeit simple) example, which can be used to illustrate nearly all the relevant kinds of algorithm, such as HMM state pruning (Lunter *et al*., 2003), Expectation Maximization (Holmes, 2005b) or alignment (Lunter *et al*., 2004).

Transducers provide a convenient bridge between rigorous phylogenetic analysis of indel processes and the rich lore of finite-state machine design. Many empirically observed characteristics of genome evolution can be integrated with transducers: they provide a systematic framework for analyzing mutation rates, including variations in GC content, fluctuating local conservation, methylation rate and codon substitution patterns (Kosiol *et al*., 2007), and for modeling phenomena involving indels, including probability distributions over exon and intron length, stop codon avoidance and conservation of codon reading frame (Kellis *et al*., 2003).

Further, transducers are not limited to models where the indel and substitution processes are independent. Extensions beyond HMM-like models allow transducers to, in principle, model microsatellite expansion/contraction, transposon insertion/deletion, local micro-duplications and micro-inversions, and various other mutation processes that would otherwise be difficult to analyze mathematically.

Formal extensions to string transducers allow them to model RNA and gene structure. Related machines, called *tree transducers* by linguists, are analogous to Pair Stochastic Context-Free Grammars and are used to analyze RNA sequences (Bradley and Holmes, 2007; Holmes, 2005a; Sakakibara, 2003).

Further discussion of transducers, including links to animations, may be found on our wiki at the following URL: http://biowiki.org/StringTransducers

2 INDEL HISTORIES

We now turn to the evolutionary model for indels described in Kim and Sinha (2007) and the associated algorithms. It can immediately be seen that the Indelign model is a transducer: conditionally normalizing the probabilities of Indelign's Pair HMM gives just the probability $P(\pi, Y|X)$ associated with the evolution $X \overset{\Delta T}{\to} Y$.

Indelign's ANNOTATE algorithm returns the maximum-likelihood indel history given a multiple alignment of n observed sequences. It operates on 'blocks', defined as the spans of maximal ungapped stretches of observed sequence, and computes the maximum-likelihood indel history of sets of consecutive dependent blocks by labeling each block as gap or non-gap for all ancestral sequences. If there are k such conditionally dependent blocks, then each node has a labeling in $\{*, -\}^k$. A dynamic-programming (DP) version of the algorithm, which iterates over combined labelings of sets of three nodes (two siblings and their parent), has a worst-case time complexity of $\mathcal{O}(N\,2^{3k})$, where N is the number of nodes in the phylogenetic tree. Note that k is theoretically bounded only by the alignment length.

This enumeration of indel histories over blocks can be contrasted with the state-enumeration approach typical of transducer DP algorithms. Alignment to a composed transducer can be expressed as a one-dimensional DP problem over an alphabet of strings in $\{A, C, G, T, -\}^N$, where each character in the string corresponds to the residue or gap at a particular node in the phylogenetic tree (with N nodes). The composed transducer has $\mathcal{O}(a^N)$ states and hence at most $\mathcal{O}(a^{2N})$ transitions, where a is the number of states of a single transducer. By analogy with standard path-inference to an HMM, we can see that this state-enumeration approach has time complexity $\mathcal{O}(L\,a^{2N})$, where L is the length of the multiple alignment. For the simplest transducers $a \simeq 3$, though it may be possible to reduce this by redundant-state elimination. In contrast to the above analysis of Indelign's ANNOTATE algorithm, this upper bound on the complexity is not input-dependent. The term 'phylo-HMM' has been coined to describe such phylogenetically-structured HMMs, particularly when the multiple alignment is supplied as an external constraint.

The $\mathcal{O}(L\,a^{2N})$ time complexity makes exact DP to a composed transducer impractical for datasets of many sequences, but Markov Chain Monte Carlo (MCMC) approaches offer a principled alternative. Given a multiple alignment of observed leaf sequences, we can sample exactly from the posterior distribution over indel histories by starting with some initial estimate of the indel history and modifying it by successive local MCMC 'moves', e.g. the branch- and node-sampling moves described in Holmes and Bruno (2001). These moves involve sampling over only local transducer compositions around the neighborhood of a single branch or node, allowing us to avoid the computational cost of inference on the full phylo-machine. Branch sampling has time complexity $\mathcal{O}(L^2\,a^2)$ and node sampling $\mathcal{O}(L\,a^6)$: much better than the $\mathcal{O}(L\,a^{2N})$ cost of exact inference.

The relative efficiency of Kim and Sinha's enumerative algorithm depends strongly on the dataset used. In Figures 1 and 2 we plot the distributions of k-values for two genomic datasets, the 12 newly sequenced *Drosophila genomes* (Drosophila Comparative Genome Sequencing and Analysis Consortium, 2007) and data from the ENCODE project (Margulies *et al*., 2007). Alignments were created with MAVID (Bray and Pachter, 2004). Most blocks of the *Drosophila* alignments belong to relatively short sequences of conditionally dependent blocks and so are amenable to analysis with Indelign, but the tail of the distribution stretches to k-values of greater than 10^4. Multiple alignments of the highly diverged genomes of the ENCODE project are dominated by very high k-values of order 10^5. Indelign's inference algorithm grows exponentially in complexity with k, making it likely

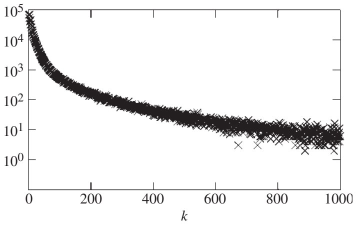

Fig. 1. Distribution of the number k of sequential conditionally dependent blocks in the MAVID alignments of the 12 *Drosophila* genomes (Drosophila Comparative Genome sequencing and Analysis Consortium, 2007). k controls the complexity of Indelign's DP algorithm as $\mathcal{O}(2^{3k})$. The frequency axis is log scale, and the k axis has been truncated at 10^3 for readability. The tail stretches to 6×10^4.

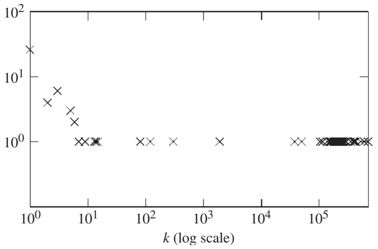

Fig. 2. Distribution of k for MAVID alignments of the ENCODE data (Margulies *et al*., 2007). Both axes are log scale.

impractical for analysis of much of this data without further heuristics or constraints.

Dynamic programming to a composed transducer, on the other hand, can handle such datasets with a complexity that is (in the worst case) exponential in the number of tree nodes, but linear in alignment length. It is even possible to achieve sub-linear memory complexity with respect to alignment length, using recursive approaches (Hirschberg, 1975; Tarnas and Hughey, 1998). Other resource-saving techniques include sparse DP algorithms such as Treeterbi (Keibler *et al*., 2007). Such time- and space-saving approaches may make analysis of even extremely long genomic sequences increasingly feasible.

Kim and Sinha note that, in practice, the actual complexity of Indelign is often significantly reduced by the restrictions on evolutionary histories that they impose, namely that (i) nucleotides cannot be deleted and then re-inserted at the same position and (ii) indel event boundaries coincide with observed gap boundaries. Both of these assumptions significantly constrain the available paths through a phylogenetically composed transducer, and so should benefit any transducer-based method. Assumption (i) is often taken as standard in the statistical alignment literature (Thorne *et al*., 1991) and is implicit in the rules for transducer composition (Holmes, 2003). Assumption (ii) can be expressed as a restriction on the transitions that the transducer can use at each particular alignment column.

A strength of Indelign's approach is the ease with which arbitrary distributions over indel lengths can be modeled. HMMs and transducers, in contrast, most naturally model geometric distributions. Extra states can be introduced to give arbitrary length distributions (this is the procedure Kim and Sinha use when describing the Pair HMM of their model) but much of the expressive power so conferred, such as long tails, can be compactly approximated by a transducer with a mixture of geometrics [see Do *et al*. (2005) for the PROBCONS program]. This is a long-understood design principle of bioinformatics state machines (Miller and Myers, 1988).

Addendum: During the review phase for this article, Diallo *et al*. (2007) published results using a phylo-HMM extremely similar to the one we have proposed in this section. In place of exact MCMC, they introduce a principled approximation that limits complexity by discarding low-valued cells from the DP matrix.

3 VERSATILE MACHINES

As we have shown, transducers provide a consistent language for many different flavors of algorithm, including multiple alignment (Hein, 2001; Holmes, 2003) and post-alignment inference (Diallo *et al*., 2007; Kim and Sinha, 2007). The theory can frame questions of computational complexity in such models.

The range of possible algorithms extends beyond maximum-likelihood inference of ancestral indel history. One can sum over histories using the Forward–Backward algorithm (Durbin *et al*., 1998; Holmes, 2003), or sample histories from the posterior distribution using various flavors of MCMC (Holmes and Bruno, 2001; Lunter *et al*., 2004). Despite several assertions in the literature that MCMC or statistical alignment are unlikely to be practical for genomes, there is no reason to anticipate that this should be so. It is possible to construct transducer-based MCMC algorithms using similar resources to pairwise alignment (Holmes, 2003; Holmes and Bruno, 2001). While unconstrained pairwise alignment of genomic-scale sequences is impractical, several methods that impose constraints to reduce memory usage can be applied to MCMC (Bray and Pachter, 2004; Metzler *et al*., 2001; Myers and Miller, 1988).

One can readily estimate evolutionary rates and other parameters for transducer models. Measurement of evolutionary rates may reveal natural selection and other interesting signatures of evolution (Holmes, 2005b; Lunter *et al*., 2006). This can be achieved either by maximum-likelihood techniques such as Expectation Maximization (Durbin *et al*.,1998; Holmes and Rubin, 2002) or by MCMC (Lunteret al., 2005; Metzler *et al*., 2001). Bayesian methods, such as the use of priors, can easily be introduced (Brown *et al*., 1993).

Many bioinformatics analyses that use multiple alignments may benefit from reformulation in terms of transducers. Examples include phylogeny, where current techniques for sampling trees can be extended to co-sample alignments (Lunter *et al*., 2005); homology profiling, where HMMs that incorporate evolution have enhanced performance (Qian and Goldstein, 2004); comparative genome annotation using phylogrammars (Klosterman *et al*., 2006); the detection of protein-coding genes via indels that preserve reading frame (Kellis *et al*., 2003) and the reconstruction of ancestral genomes (Ma *et al*., 2006). Transducers can also be used to model local context-dependent mutations, such as simultaneous substitutions at adjacent nucleotides (Averof *et al*., 2000), other context-dependent substitutions such as CpG effects (Lunter and Hein, 2004; Siepel and Haussler, 2004), or expansion/contraction of microsatellites. Tree transducers (Fülöp and

Vogler, 1998) can be used to model the evolution of structured features such as non-coding RNA genes (Holmes, 2005a) or protein-coding genes (Carmel *et al.*, 2005). It may even be possible to model more context-dependent mutations, such as local duplications, inversions or transpositions, using models related to transducers.

3.1 Transducer software and algorithms

Several software tools for working with transducers are in common circulation, some of them unpublished. Tools for statistical alignment, phylogeny and/or parameter estimation include Handel (Holmes and Bruno, 2001); Phylogeny Café (Miklos *et al.*, 2007); BEAST (Drummond and Rambaut, 2003); BAli-Phy (Suchard and Redelings, 2006); MCMCALGN (Fleissner *et al.*, 2005; Metzler *et al.*, 2001) MCALIGN (Wang *et al.*, 2006), PRANK (Loytynoja and Goldman, 2005) and Indelign (Kim and Sinha, 2007). Our lab provides several transducer-related tools and resources, including short illustrative animations (biowiki.org/PhyloFilm).

ACKNOWLEDGEMENTS

I.H. was funded in part by NIH/NHGRI grant 1R01GM076705-01. RB was supported under an NSF Graduate Research Fellowship.

Conflict of Interest: none declared.

REFERENCES

Averof,M. *et al.* (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, **287**, 1283–1286.

Bishop,M.J. and Thompson,E.A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.*, **190**, 159–165.

Bradley,R.K. and Holmes,I. (2007) RNA structure evolution and transducer composition, in preparation.

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

Brown,M. *et al.* (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter,L. *et al.* (eds.), *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 47–55.

Carmel,L. *et al.* (2005) An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes. In *Lecture Notes in Bioinformatics 3678: Proceedings of RECOMB 2005 Comparative Genomics International Workshop (RCG 2005)*. vol. 3678. Springer, pp. 35–46.

Churchill,G.A. (1992) Hidden markov chains and the analysis of genome structure. *Comput. Chem.*, **16**, 107–115.

Diallo,A.B *et al.* (1992) Exact and heuristic algorithms for the indel maximum likelihood problem. *J. Comput. Bio.*, **14**, 446–461.

Do,C.B. *et al.* (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340. Comparative Study.

Drosophila Comparative Genome Sequencing and Analysis Consortium (2007) Evolution of genes and genomes in the genus drosophila, in preparation.

Drummond,A.J. and Rambaut,A. (2003) BEAST v1.0. Available from http://evolve.zoo.ox.ac.uk/beast/

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Fleissner,R. *et al.* (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol.*, **54**, 548–561. Comparative Study.

Fülöp,Z. and Vogler,H. (1998) *Syntax-Directed Semantics: Formal Models Based on Tree Transducers*, Springer.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Hein,J. (2002) An algorithm for statistical alignment of sequences related by a binary tree. In Altman,R.B. *et al.* (eds.), *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp.179–190.

Hirschberg,D.S. (1975) A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, **18**, 341–343.

Holmes,I. (2003) Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, **19** (Suppl. 1), i147–i157.

Holmes,I. (2005a) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**.

Holmes,I. (2005b) Using evolutionary Expectation Maximization to estimate indel rates. *Bioinformatics*, **21**, 2294–2300.

Holmes,I. and Bruno,W.J. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**, 803–820.

Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *J. Comput. Biol.*, **5**, 493–504.

Holmes,I. and Rubin,G.M. (2002) An Expectation Maximization algorithm for training hidden substitution models. *J. Mol. Bio.*, **317**, 757–768.

Keibler,E. *et al.* (2007) The Treeterbi and Parallel Treeterbi algorithms: efficient, optimal decoding for ordinary, generalized, and Pair HMMs. *Bioinformatics*.

Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

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

Klosterman,P.S. *et al.* (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, **7**.

Knudsen,B. and Miyamoto,M. (2003) Sequence alignments and pair hidden Markov models using evolutionary history. *J. Mol. Biol.*, **333**, 453–460.

Kosiol,C. *et al.* (2007) An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.*

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

Lunter,G. and Hein,J. (2004) A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, **20** (Suppl. 1), I216–I223.

Lunter,G. *et al.* (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83.

Lunter,G. *et al.* (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.*, **2**.

Lunter,G.A. *et al.* (2003) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comput. Biol.*, **10**, 869–889.

Lunter,G.A. *et al.* (2004) Statistical alignment: recent progress, new applications, and challenges. In Nielsen,R. (ed.) *Statistical Methods in Molecular Evolution. Series in Statistics in Health and Medicine*. Springer Verlag.

Ma,J. *et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, **16**, 1557–1565. Comparative Study.

Margulies,E. *et al.* (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.*, in press.

Mealy,G.H. (1955) A method for synthesizing sequential circuits. *Bell Syst. Tech. J.*, **34**, 1045–1079.

Metzler,D. *et al.* (2001) Assessing variability by joint sampling of alignments and mutation rates. *J. Mol. Evol.*, **53**, 660–669.

Miklós,I. *et al.* (2004) A long indel model for evolutionary sequence alignment. *Mol. Biol. Evol.*, **21**, 529–540.

Miklós,I *et al.* (2007) Phylogeny Cafe. Available from http://phylogeny-cafe.elte.hu/

Miller,W. and Myers,E.W. (1988) Sequence comparison with concave weighting functions. *Bull. Math. Biol.*, **50**, 97–120.

Mohri,M. *et al.* (2000) Weighted finite-state transducers in speech recognition. *ISCA ITRW Automatic Speech Recognition*, pp. 97–106.

Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci.*, **4**, 11–17.

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.

Qian,B. and Goldstein,R.A. (2004) Performance of an iterated T-HMM for homology detection. *Bioinformatics*.

Sakakibara,Y. (2003) Pair Hidden Markov Models on Tree Structures. Evaluation Studies, pp. 232–240.

Searls,D.B. and Murphy,K.P. (1995) Automata-theoretic models of mutation and alignment. In Rawlings,C. *et al.* (eds.), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 341–349.

Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, **21**, 468–488.

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

Suchard,M.A. and Redelings,B.D. (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**, 2047–2048.

Tarnas,C. and Hughey,R. (1998) Reduced space hidden Markov model training. *Bioinformatics*, **14**, 401–406.

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

Wang,J. *et al.* (2006) MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics*, **7**, 292.