BMC Bioinformatics

Methodology article

Bayesian coestimation of phylogeny and sequence alignment

Gerton Lunter*[1], István Miklós[2], Alexei Drummond[3], Jens Ledet Jensen[4] and Jotun Hein[1]

Address: [1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK, [2]MTA-ELTE Theoretical Biology and Ecology Group, Pázmány Péter sétány 1/c 1117 Budapest, Hungary, [3]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK and [4]Department of Mathematical Sciences, University of Aarhus, Ny Munkegade, Building 530, DK-8000 Aarhus C, Denmark

Email: Gerton Lunter* - lunter@stats.ox.ac.uk; István Miklós - miklosi@ramet.elte.hu; Alexei Drummond - alexei.drummond@zoology.oxford.ac.uk; Jens Ledet Jensen - jlj@imf.au.dk; Jotun Hein - hein@stats.ox.ac.uk

* Corresponding author

Published: 01 April 2005

BMC Bioinformatics 2005, **6**:83 doi:10.1186/1471-2105-6-83

Received: 20 January 2005
Accepted: 01 April 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/83

Abstract

Background: Two central problems in computational biology are the determination of the alignment and phylogeny of a set of biological sequences. The traditional approach to this problem is to first build a multiple alignment of these sequences, followed by a phylogenetic reconstruction step based on this multiple alignment. However, alignment and phylogenetic inference are fundamentally interdependent, and ignoring this fact leads to biased and overconfident estimations. Whether the main interest be in sequence alignment or phylogeny, a major goal of computational biology is the co-estimation of both.

Results: We developed a fully Bayesian Markov chain Monte Carlo method for coestimating phylogeny and sequence alignment, under the Thorne-Kishino-Felsenstein model of substitution and single nucleotide insertion-deletion (indel) events. In our earlier work, we introduced a novel and efficient algorithm, termed the "indel peeling algorithm", which includes indels as phylogenetically informative evolutionary events, and resembles Felsenstein's peeling algorithm for substitutions on a phylogenetic tree. For a fixed alignment, our extension analytically integrates out both substitution and indel events within a proper statistical model, without the need for data augmentation at internal tree nodes, allowing for efficient sampling of tree topologies and edge lengths. To additionally sample multiple alignments, we here introduce an efficient partial Metropolized independence sampler for alignments, and combine these two algorithms into a fully Bayesian co-estimation procedure for the alignment and phylogeny problem.

Our approach results in estimates for the posterior distribution of evolutionary rate parameters, for the maximum *a-posteriori* (MAP) phylogenetic tree, and for the posterior decoding alignment. Estimates for the evolutionary tree and multiple alignment are augmented with confidence estimates for each node height and alignment column. Our results indicate that the patterns in reliability broadly correspond to structural features of the proteins, and thus provides biologically meaningful information which is not existent in the usual point-estimate of the alignment. Our methods can handle input data of moderate size (10–20 protein sequences, each 100–200 bp), which we analyzed overnight on a standard 2 GHz personal computer.

Conclusion: Joint analysis of multiple sequence alignment, evolutionary trees and additional evolutionary parameters can be now done within a single coherent statistical framework.

Background

Two central problems in computational biology are the determination of the alignment and phylogeny of a set of biological sequences. Current methods first align the sequences, and then infer the phylogeny given this fixed alignment. Several software packages are available that deal with one or both of these sub-problems. For example, ClustalW [1] and T-Coffee [2] are popular sequence alignment packages, while MrBayes [3], PAUP* [4] and Phylip [5] all provide phylogenetic reconstruction and inference. Despite working very well in practice, these methods share some problems. First, the separation into a multiple-alignment step and a phylogenetic inference step, is fundamentally flawed. The two inference problems are mutually dependent, and alignments and phylogeny should ideally be co-estimated, a point first made by Sankoff, Morel and Cedergren [6]. Indeed, a proper weighting of mutation events in multiple sequences requires a tree, which in turn can only be determined if a multiple alignment is available. For instance, ClustalW and T-Coffee compute their alignments based on a neighbour-joining guide tree, biasing subsequent phylogenetic estimates based on the resulting alignment. Moreover, fixing the alignment after the first step ignores the residual uncertainty in the alignment, resulting in an overconfident phylogenetic estimate.

This leads on to the second issue, which is that heuristic methods are used to deal with insertions and deletions (indels), and sometimes also substitutions. This lack of a proper statistical framework makes it very difficult to accurately assess the reliability of the alignment estimate, and the phylogeny depending on it.

The relevance of statistical approaches to evolutionary inference has long been recognised. Time-continuous Markov models for substitution processes were introduced more than three decades ago [7]. Inference methods based on these have been considerably improved since then [8], and now have all but replaced older parsimony methods for phylogeny reconstruction. With alignments, progress towards statistically grounded methods has been slower. The idea to investigate insertions and deletions in a statistical framework was first considered by Bishop and Thompson [9]. The first evolutionary model, termed the TKF91 model, and corresponding statistical tools for pairwise sequence alignment were published by Thorne, Kishino and Felsenstein [10]. Its extension to multiple sequences related by a tree has been intensively investigated in the last few years [11-17], and has recently also been extended to RNA gene evolution [18]. Current methods for statistical multiple alignment often computationally demanding, and full maximum likelihood approaches are limited to small trees. Markov chain Monte Carlo techniques can extend these methods to practical problem sizes.

Statistical modelling and MCMC approaches have a long history in population genetic analysis. In particular, coalescent approaches to genealogical inference have been very successful, both in maximum likelihood [19,20] and Bayesian MCMC frameworks [21,22]. The MCMC approach is especially promising, as it allows the analysis of large data sets, as well as nontrivial model extensions, see e.g. [23]. Since divergence times in population genetics are small, alignment is generally straightforward, and genealogical inference from a fixed alignment is well-understood [20,24-26]. However, these approaches have difficulty dealing with indels when sequences are hard to align. Indel events are generally treated as missing data [27], which renders them phylogenetically uninformative. This is unfortunate as indel events can be highly informative of the phylogeny, because of their relative rarity compared to substitution events. Statistical models of alignment and phylogeny often refer to missing data. Not all of these can be integrated out analytically (e.g. tree topology), and these are dealt with using Monte Carlo methods. The efficiency of such approaches depend to a great extent on the choice of missing data. In previous approaches to statistical alignment, the sampled missing data were either unobserved sequences at internal nodes [28], or both internal sequences and alignments between nodes [13], or dealt exclusively with pairwise alignments [29,30]. In all cases the underlying tree was fixed. In [31] we published an efficient algorithm for computing the likelihood of a multiple sequence alignment under the TKF91 model, given a fixed underlying tree. The method analytically sums out all missing data (pertaining to the evolutionary history that generated the alignment), eliminating the need for any data augmentation of the tree. This methodology is referred to in the MCMC literature as *Rao-Blackwellization* [32]. As a result, we can treat indels in a statistically consistent manner with no more than a constant multiplicative cost over existing methods that ignore indels.

The only missing ingredient for a full co-estimation procedure is an alignment sampler. Unfortunately, there exists no Gibbs alignment sampler that corresponds to the analytic algorithm referred to above. In this paper we introduce a partial importance sampler to resample alignments, based on a proposal mechanism built on a partial score-based alignment procedure. This type of sampler supports the data format we need for efficient likelihood calculations, while still achieving good mixing in reasonable running time (see Results).

We implemented the likelihood calculator and the alignment sampler in Java, and interfaced them with an

existing MCMC kernel for phylogenetics and population genetics [22]. We demonstrate the practicality of our approach on an analysis of 10 globin sequences.

Results

Definition of the TKF model

The TKF91 model is a continuous-time reversible Markov model describing the evolution of nucleotide (or amino acid) sequences. It models three of the main processes in sequence evolution, namely *substitutions*, *insertions* and *deletions* of characters, approximating these as single-character processes. A sequence is represented as a string alternatingly consisting of *links* and characters connected by these links. This string both starts and terminates with a link. Insertions and deletions are modeled through a time-continuous birth-death process of links. When a new link is born, its associated character (by convention, its right neighbour) is chosen from the equilibrium distribution of the substitution process. (The original TKF91 model used a simple substitution process, the Felsenstein-81 model [27]. It is straightforward to replace this by more general nucleotide or amino acid substitution models [33].) When a link dies, its associated character dies too. The leftmost link of the sequence has no corresponding character to its left, and is never deleted. For this reason it is called the *immortal link*.

Since subsequences evolve independently, it is sufficient to describe the evolution of a single character-link pair. In a given finite time span, this pair evolves into a finite subsequence of characters and links. Since insertions originate from links, only the first character of this descendant subsequence may be homologous to the original character, while subsequent ones will have been inserted and therefore not be homologous to ancestral characters. The model as applied to pairwise alignments was solved analytically in [10], see also [34]. Conceptually, the model can be trivially extended to trees, but the corresponding algorithms for likelihood calculations have been developed only recently [11,12,14-16].

Because the TKF91 model is time reversible, the root placement does not influence the likelihood, an observation known as Felsenstein's "Pulley Principle" [27]). Although the algorithms we developed are not manifestly invariant under changes in root placement, in fact they are. We have used time reversibility to check correctness of our implementations.

Computing the likelihood of a homology structure

The concept of *homology structure* [31], also known as effective alignment [35], refers to an alignment of sequences at leaves without reference to the internal tree structure, and without specifying the ordering of exchangable columns (see below for more details). We derived a linear-time algorithm that computes the likelihood of observing a set of sequences and their homology structure, given a phylogeny and evolutionary parameters, under the TKF91 model [31]. By definition, this likelihood is the sum of the probabilities of all evolutionary scenarios resulting in the observed data. It was previously shown that such evolutionary scenarios can be described as a path in a multiple-HMM ([13,28]), and the likelihood can thus be calculated as the sum of path probabilities over all such paths, in time polynomial in the number of states. However, this straightforward calculation is infeasible for practical-sized biological problems, since the number of states in the HMM grows exponentially with the number of sequences [16]. Since our algorithm does not feature this exponential blow-up of Markov states, we termed it the *one-state recursion*. In contrast to previous approaches [13,28], the one-state recursion relieves us from the need to store missing data at internal tree nodes, allowing us to change the tree topology without having to resample this missing data. This enables us to consider the *tree* as a parameter, and efficiently sample from tree space. The concept of homology structure referred to above is key to our algorithm, and we will presently define this concept more precisely. Let A_1, A_2, ...A_m be sequences, related by a tree T with vertex set V. Let a_i^j denote the jth character of sequence A_i, and let A_i^k denote its k long prefix. A *homology structure* $\mathcal{H}$ on A_1, ..., A_m is an equivalence relation $\sim$ on the set of all the characters of the sequences, $C = \{ a_i^j \}$, specifying which characters are homologous to which. The evolutionary indel process generating the homology structure on the sequences imposes constraints on the equivalence relations that may occur. More precisely, the equivalence relation $\sim$ has the property that a total ordering, $<_h$, exists on C such that

$$\begin{aligned} a_i^p =_h a_j^q \quad &\Leftrightarrow \quad a_i^p \sim a_j^q, \\ a_i^p <_h a_i^q \quad &\Leftrightarrow \quad p < q \end{aligned} \qquad (1)$$

(Here, $a =_h b$ is equivalent to: $a \not<_h b$ and $b \not<_h a$.) In particular, these conditions imply that the characters constituting a single sequence are mutually nonhomologous. The ordering $<_h$ corresponds to the ordering of columns of homologous characters in an alignment. Note that for a given homology structure, this ordering may not be unique (see Fig. 1). This many-to-one relationship of alignment to homology structure is the reason for introducing the concept of homology structure, instead of using the more common concept of alignment.

The one-state recursion, which calculates the likelihood of a homology structure, is a convolution of two dynamic

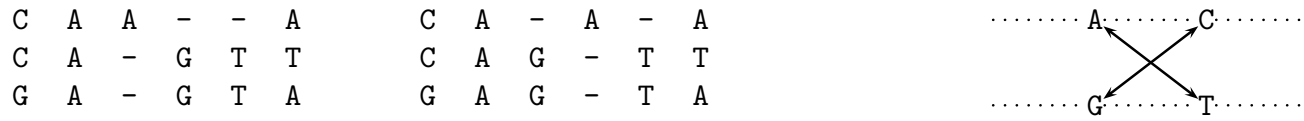

Figure 1
Alignments and homology structure. (Left:) Two alignments representing the same homology structure. A "homology structure" is defined as the set of all homology relationships between residues from the given sequences; residues are homologous if they appear in the same alignment column. Our recursion includes contributions from all alignments compatible with a given homology structure (itself represented by a single alignment). (Right:) Due to the evolutionary process acting on the sequences, homology relationships (arrows) will never 'cross' as depicted. This restriction on the equivalence relation ~ is codified by $<_h$ (see text).

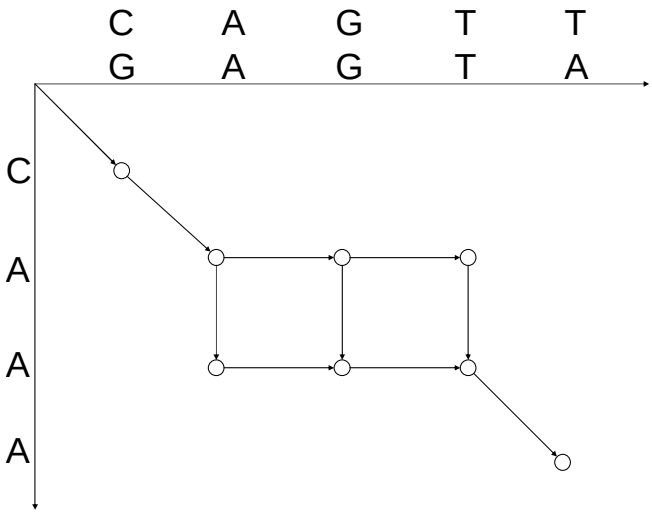

Figure 2
Dynamic programming table traversal. The multiple alignment prefixes (represented by o symbols) traversed by the one-state recursion, when the input is the homology structure of Fig. 1. (For clarity, the vectors are plotted in two dimensions instead of the actual three.) The homology structure is represented by the graph, and each directed path on this graph uniquely corresponds to an alignment that is compatible with the homology structure.

programming algorithms. The top-level algorithm traverses the prefix set of the multiple alignments representing the homology structure (see Figure 2). This repeatedly calls on a reverse traversal algorithm on the phylogenetic tree, which sums out the likelihood contributions of substitutions and indels under the TKF91 model. See [31] for full details.

A partial Metropolized independence sampler

Because our algorithm does not require the phylogenetic tree to be augmented with missing data, proposing changes to the evolutionary tree is easy, and mixing in tree space is very good. The drawback however is that without data augmentation, it is unclear how to perform Gibbs sampling of alignments, and we have to resort to other sampling schemes. One straightforward choice would be a standard Metropolis-Hastings procedure with random changes to the alignment, but we expect slow mixing from such an approach. Another general approach is Metropolized independence sampling. Its performance depends on the difference between the proposal distribution and the target distribution, and this will inevitably become appreciable with growing dimension of the problem, as measured by the number and length of the sequences to be aligned. We therefore opted for a *partial Metropolized independence sampler* [36], where we partly defy the "curse of dimensionality" by resampling only a segment of the current alignment. Above increasing the acceptance ratio, this method has the added advantage of being a more efficient proposal scheme, since the time complexity of the algorithm is proportional to the square of the window size, and so leads to an effective increase in mixing per processor cycle. Metzler *et al*. [29] followed a parallel approach, using a partial Gibbs sampler, and showed that this resulted in faster mixing compared to a full Gibbs sampling step. Since the realignment step may change the window length (measured in alignment columns), to have a reversible Markov chain we need all window sizes to have positive proposal probability. We chose a geometric length distribution, but other distributions can be considered equally well.

The proposal algorithm

The proposal algorithm is as follows. A window size and location is proposed, the alignment of subsequences within this window is removed, and a new alignment is proposed by a stochastic version of the standard score-based progressive pairwise alignment method. First, dynamic programming (DP) tables are filled as for a deterministic score-based multiple alignment, starting at the tree tips and working towards the root, aligning sequences

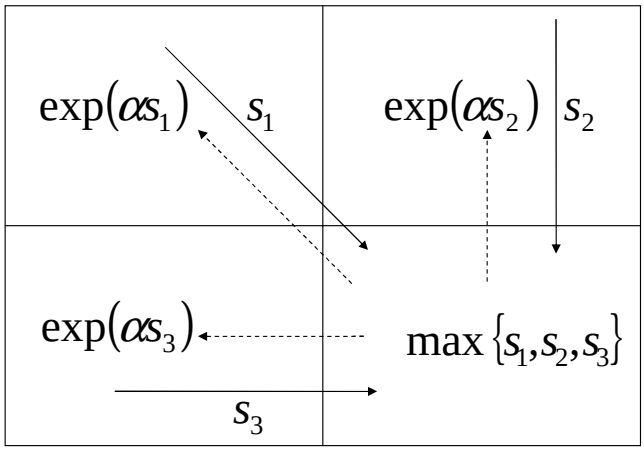

Figure 3
Generating the proposal alignment. This figure illustrates the stochastic sequence aligner. In the deterministic fill-in process, the three scores are s_1, s_2 and s_3, hence the value in this cell is $\max\{s_1, s_2, s_3\}$. In the stochastic traceback phase, the three neighbor cells are choosen with probabilities proportional to $\exp(\alpha s_i)$, where $\alpha > 0$ is a scaling parameter. The chosen traceback path corresponds to the proposed alignment in the usual way.

and profiles. We used linear gap penalties, and a similarity scoring matrix that was obtained by taking the log-odds of a probabilistic substitution matrix. The underlying phylogeny was used to define divergence times, and served as alignment guide tree. After filling the DP tables, we applied stochastic traceback. The probabilities for the three possible directions at each position was taken to be proportional to $\exp(\alpha s)$, where s is the deterministic score and α is a scale parameter (see Fig. 3). The set of paths that emerged in this way then determined the multiple alignment. All possible alignments can be proposed in this manner, and the proposal as well as the back-proposal probabilities can be calculated straightforwardly.

Correctness of the sampler

There are two problems with the proposal sampler introduced above. First, we propose alignments instead of homology structures. We need the latter, since the algorithm derived in this paper calculates the likelihood of the homology structure, not the particular alignment. Although it would be conceptually and (for the sampler) computationally simpler to use alignments, we are not aware of any efficient algorithm that can calculate such *alignment* likelihoods. The second problem is that calculating the proposal probability of a particular alignment is not straightforward. Any choice of window size and location may result in the same proposal alignment. To calculate the true proposal probability of particular alignments, we need to sum over all possible windows, which is prohibitively expensive.

Fortunately, we can solve both problems efficiently. We can sample alignments uniformly inside a homology structure, and at the same time sample homology structures according to their posterior probabilities. As biologically meaningful questions refer to homologies and not particular alignments, it seems reasonable to impose a simple uniform distribution over alignments within homology structures. The second problem is solved by not calculating an alignment proposal probability, but the proposal probability of the combination of an alignment and a resampling window. For a proposal of alignment X_2 and window w from a current alignment X_1, we use the following Metropolis-Hastings ratio:

$$\min\left\{1, \frac{|H_1| \times \pi(H_2) \times T(X_1, w \mid X_2)}{|H_2| \times \pi(H_1) \times T(X_2, w \mid X_1)}\right\}, \qquad (2)$$

where H_1 and H_2 are homology structures corresponding to the alignments X_1 and X_2 respectively, $|H_1|$ and $|H_2|$ are their cardinalities (i.e. the number of alignments representing these homology structures), and T is the proposal probability. Using this ratio, the Markov chain will converge to the desired distribution $\pi(X) = \pi(H)/|H|$, since the detailed balance condition is satisfied. Indeed,

$$\begin{aligned}\pi(X_1)P(X_2 \mid X_1) &= \frac{\pi(H_1)}{|H_1|}\sum_{w \in W} T(X_2, w \mid X_1)\min\left\{1, \frac{|H_1|\pi(H_2)T(X_1, w \mid X_2)}{|H_2|\pi(H_1)T(X_2, w \mid X_1)}\right\} \\ &= \sum_{w \in W}\min\left\{\frac{\pi(H_1)T(X_2, w \mid X_1)}{|H_1|}, \frac{\pi(H_2)T(X_1, \boldsymbol{w} \mid X_2)}{|H_2|}\right\} \\ &= \pi(X_2)P(X_1 \mid X_2), \end{aligned} \qquad (3)$$

where the final equality holds because of the symmetry of the left-hand side. The cardinality of a homology structure, $|H_1|$, is the number of possible directed paths in the graph spanned by the one-state recursion; in other words, the number of permutations of alignment columns that result in alignments compatible with the given homology structure (see Fig. 2). This number can be calculated straightforwardly using a dynamic programming algorithm that traverses the one-state recursion graph [31,37].

Discussion

The one-state recursion provides a method for calculating the likelihood $L = \Pr\{A, \mathcal{H} \mid T, Q, \lambda, \mu\}$ of observing the sequences with their homology structure (loosely, "alignment") given the tree and model parameters. Here A are the amino acid sequences, $\mathcal{H}$ is their homology structure, T is the tree including branch lengths, Q is the substitution rate matrix, and λ, μ are the amino acid insertion and deletion rates. To demonstrate the practicality of the new algorithm for likelihood calculation we undertook a Bayesian MCMC analysis of ten globin protein sequences (see Additional file: 1). We chose to use the

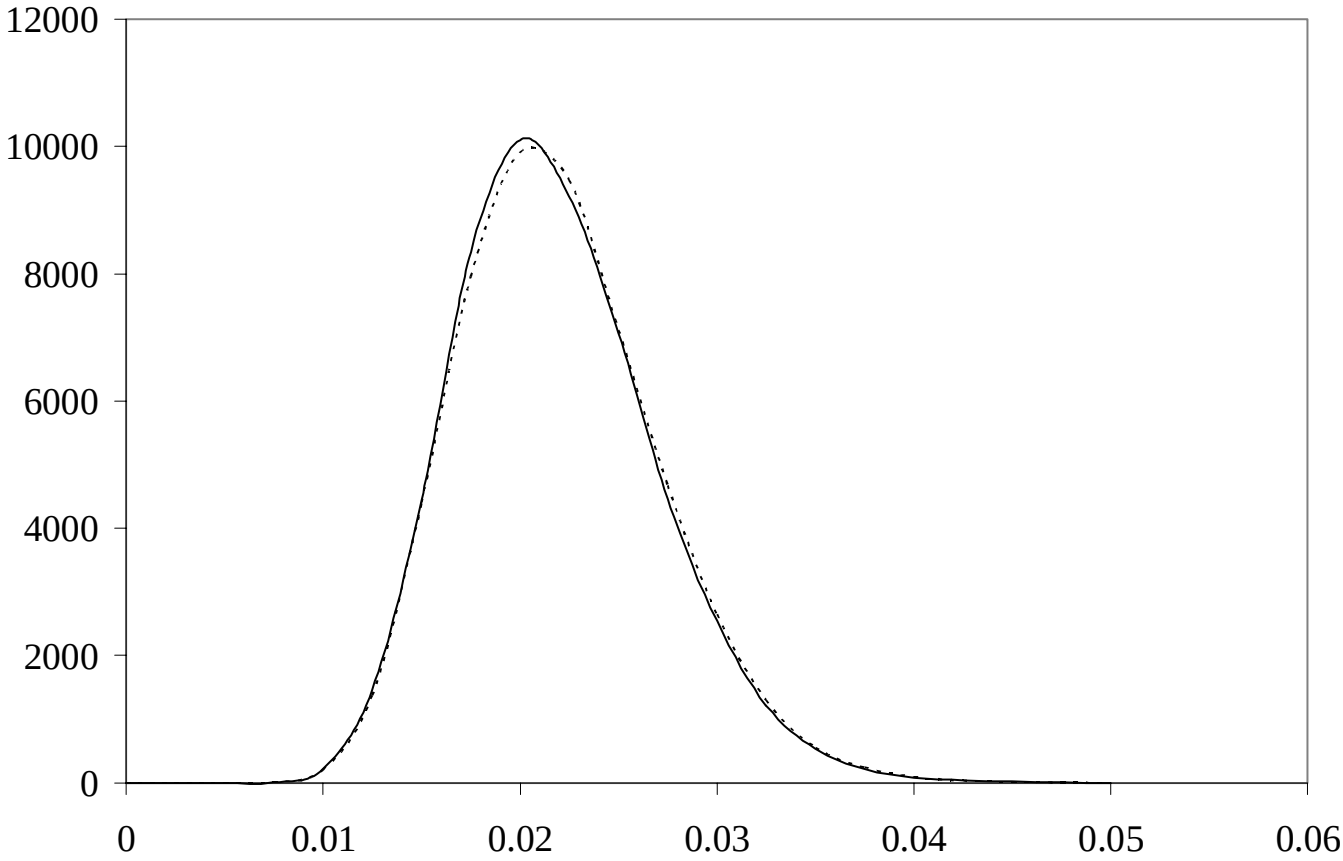

Figure 4
Posterior distribution of deletion rate μ. Estimated posterior densities of the deletion rate μ sampled according to h (see text), for two independent runs, suggesting excellent convergence. The sampled mean is 0.0207; the 95% highest posterior density (HPD) interval was estimated to be (0.0121, 0.0316).

standard Dayhoff rate matrix to describe the substitution of amino acids. As initial homology structure we used the alignment computed by T-Coffee. We co-estimated homology structures, the parameters of the TKF91 model, and the tree topology and branch lengths. To do this we sampled from the posterior,

$$h(\mu, T, \mathcal{H}) = \frac{1}{Z} \Pr\{A, \mathcal{H} | T, Q, \lambda, \mu\} f(T, \lambda, \mu), \quad (4)$$

where Z is the unknown normalising constant. We chose the prior distribution on our parameters, $f(T, \lambda, \mu)$, so that T was constrained to a molecular clock, and $\lambda = \mu L/(L + 1)$ to make the expected sequence length under the TKF91 model agree with the observed lengths; here L is the geometric average sequence length. All other parameters were sampled under uniform priors. We assume a molecular clock to gain insight into the relative divergence times of the alpha-, beta- and myoglobin families. In doing so we incorporate insertion-deletion events as informative events in the evolutionary analysis of the globin family. The posterior density h is a complicated function defined on a space of high dimension. We summarise the information it contains by computing the expectations, over h, of various statistics of interest. We estimate these expectations by using MCMC to sample from h. Marginalizations for continuous variables can be done in a straightforward manner; see for example Figure 4, which depicts the marginal posterior density of the μ parameter for two independent MCMC runs, showing excellent convergence.

For alignments, the maximum *a-posteriori* alignment is very hard to estimate from an MCMC sample run, as there are typically far too many plausible alignments contributing to the likelihood. Indeed, we found that almost all alignments in a moderately long MCMC run (50000 samples) were unique. However, it is possible to reconstruct a reliable *maximum posterior decoding* [38] alignment from such a moderate long sampling run. This alignment uses

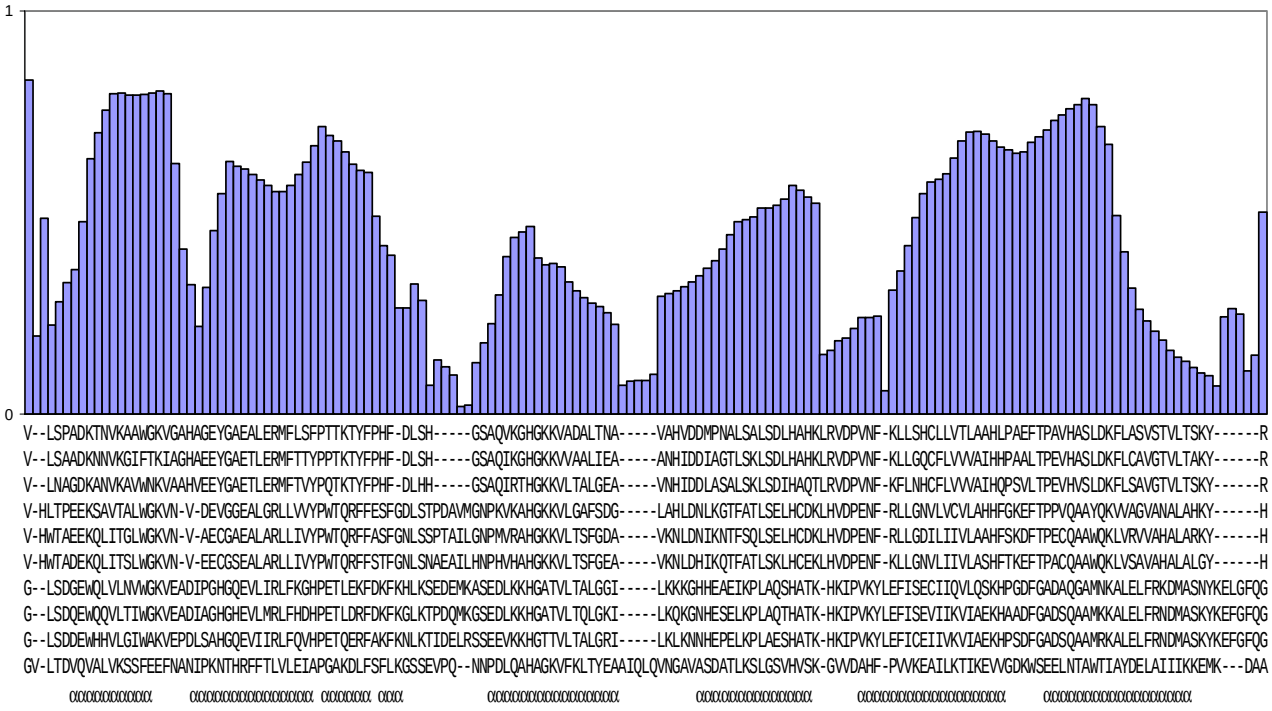

Figure 5

Maximum posterior decoding alignment, and column reliabilities. The maximum posterior decoding alignment of ten globins (human, chicken and turtle alpha hemoglobin, beta hemoglobin, myoglobin and bean leghemoglobin). Posterior probabilities for aligned columns were estimated as their rate in the Markov chain. Common alpha helices are indicated with 'α' symbols under the alignment. A broad correspondence between peaks in the posterior alignment reliability and the position of conserved secondary structure is apparent.

the posterior single-column probabilities, which can be estimated much more reliably since many alignments share particular columns, to obtain an alignment that maximizes the product of individual column posteriors. This alignment can be obtained by a simple dynamic programming algorithm [39], see Fig. 5. It is hard to visualise alternative suboptimal alignments, but the individual posterior column probabilities clearly reveal the more and less reliable regions of the alignment. We found that the reliable alignment regions broadly correspond to the alpha helical structure of the globin sequences.

Figure 6 depicts the maximum *a posteriori* (MAP) estimate of the phylogenetic relationships of the sequences. This example exhibits only limited uncertainty in the tree topology, however we observed an increased uncertainty for trees that included divergent sequences, such as bacterial and insect globins (results not shown).

The estimated time of the most recent common ancestor of each of the alpha, beta and myoglobin families are all mutually compatible (result not shown), suggesting that the molecular clock hypothesis is at least approximately valid. Analysis of a four sequence dataset demonstrate consistency in μ estimates between MCMC and previous ML analyses [16] (data not shown). Interestingly, the current larger dataset supports a lower value of μ. This is probably due to the fact that no indels are apparent within any of the subfamilies despite a considerable sequence divergence. The indel rate estimated by the current cosampling procedure is greater than the estimate on a fixed multiple alignment [31] (0.0207 vs. 0.0187), but this discrepancy is not significant for the current dataset. It should be stressed that the two MCMC analyses of the globin data set presented here are purely illustrative of the practicality of the algorithm described, and no novel biological results were obtained. The two MCMC runs of 5 million states each required less than 12 hours of CPU

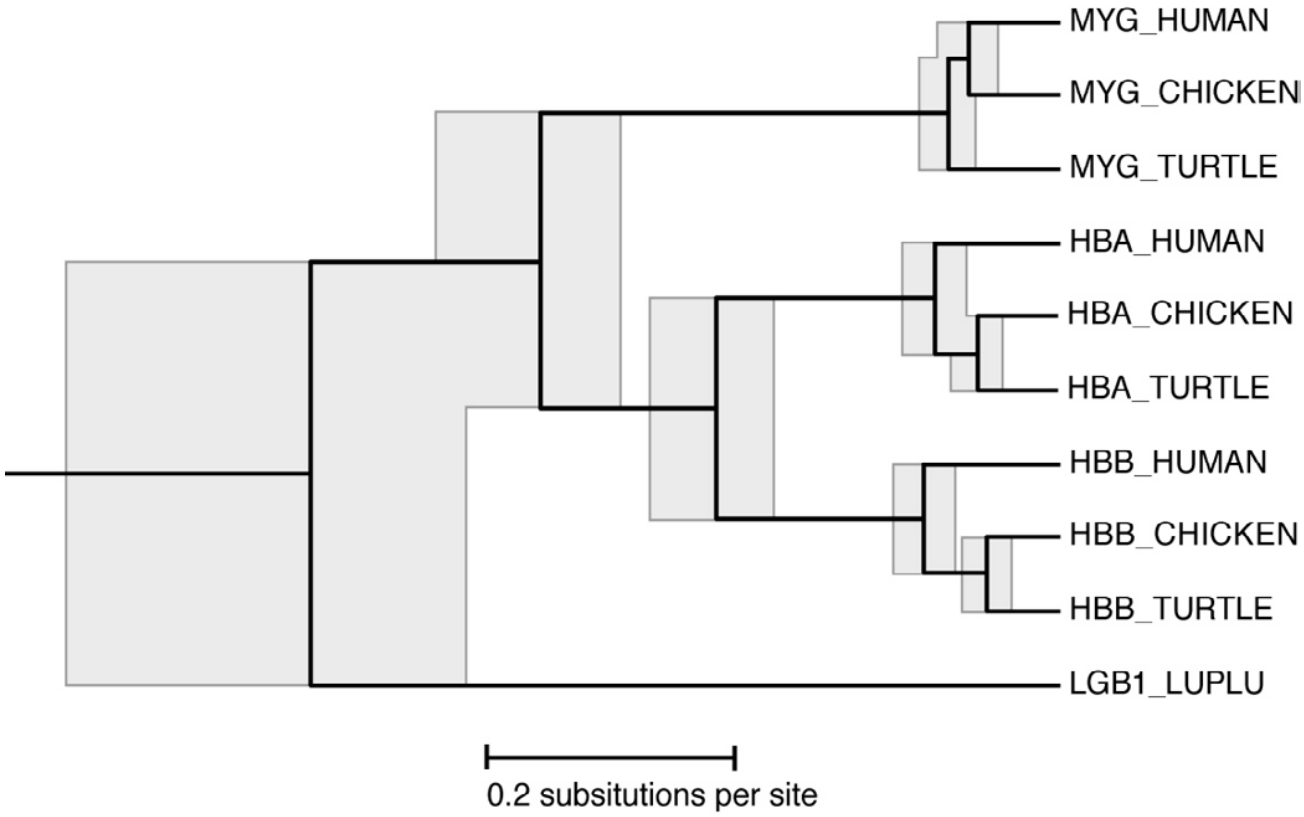

Figure 6
Maximum *a-posteriori* phylogeny. The maximum *a posteriori* tree (black) relating the ten globins of Fig. 5, and 95% confidence intervals of the node heights (grey boxes). Most of the tree's topology is well determined, with the exception of the myoglobin sub-tree. Alpha and beta chain sub-families both support the traditional ordering of birds, turtles and mammals, while the three myoglobin sequences support an unconventional phylogeny, as previously observed by Hedges and Poling [41]. However, the posterior probability for the topology of the myoglobin subtree is smaller than that for the remaining topology. The marginal posterior probability (estimated from the MCMC chain) for the monophyly of human and chicken myoglobin is 83.1%, followed by the conventional grouping of turtle and chicken at 11.9%. The third topologlcal arrangement of myoglobin occurred the remaining 5% of the time, suggesting significant homoplasy in this sub-family.

time each on a 2.0 GHz G5 Apple Macintosh running OS X, using an unoptimised implementation of the algorithm. From these runs we sampled 50000 states each. The estimated number of independent samples (estimated sample size, ESS) for the posterior probabilities was 250 and 240, respectively (see [22] for methods), while for the indel rate μ the ESSs were calculated at 5400 and 4000. We expect analyses of data sets of around 50 sequences to be readily attainable with only a few days computation.

Conclusion

In this paper we present a new cosampling procedure for phylogenetic trees and sequence alignments. The underlying likelihood engine uses recently introduced and highly efficient algorithms based on an evolutionary model (the Thorne-Kishino-Felsenstein model) that combines both the substitution and insertion-deletion processes in a principled way [31]. We show that the proposed method is applicable to medium-sized practical multiple alignment and phylogenetic inference problems.

One motivation for using a fully probabilistic model, and for using a co-estimation procedure for alignments and phylogeny, is that this makes it possible to assess the uncertainties in the inferences. Fixing either the alignment or the phylogeny leads to an underestimate of the uncertainty in the other, and score-based methods give no assessment of uncertainty whatsoever.

We show that the confidence estimates so obtained can contain biologically meaningful information. In the case of the multiple alignment of globin sequences, peaks in the posterior column reliabilities correspond broadly to the various conserved alpha helices that constitute the sequences (see Fig. 5). In the case of the tree estimate, the non-traditional phylogeny supported by the myoglobin subtree coincided with a significant polyphyly, as indicated by the posterior tree topology probabilities, and graphically represented by significantly overlapping 95% node height confidence boxes (see Fig. 6). It is clear that such confidence information significantly contributes to the usefulness of the inference.

At the heart of the method lies a recently introduced algorithm, termed the "indel peeling algorithm", that extends Felsenstein's peeling algorithm to incorporate insertion and deletion events under the TKF91 model [31]. This renders indel events informative for phylogenetic inference. Although incurring considerable algorithmic complications, the resulting algorithm is still linear-time for biological alignments (see also Figure 1). Moreover, our approach allows efficient sampling of tree topologies, as no data is presented at internal nodes.

We also developed a method for sampling multiple alignments, which is applicable for the data augmentation scheme we used for the efficient likelihood calculations. By combining the two samplers, we can co-sample alignments, evolutionary trees and other evolutionary parameters such as indel and substitution rates. The resulting samples from the posterior distribution can be summarized in traditional ways. We obtained maximum *a-posteriori* estimates of alignment, tree and parameters, and augmented these with estimates of reliability.

As was already mentioned in [10], it would be desirable to have a statistical sequence evolution model that deals with 'long' insertions and deletions, instead of single nucleotides at a time. For score-based algorithms, this is analogous to the contrast between linear and affine gap penalties. It is clear that the extension of the model to include long indels would result in considerable improvements, but the algorithmic complexities are considerable. We have made progress on a full likelihood method for statistical sequence alignment under such an evolutionary model [17], but the generalization of this method seems nontrivial. We believe that here too, Markov chain Monte Carlo approaches, combined with data augmentation, will be essential for practical algorithms. However, we also believe that in certain restricted but biologically meaningful situations, such as highly conserved proteins, the TKF91 model is reasonably realistic for the co-estimation procedure presented here to be of practical interest.

Availability and requirements

The BEAST package (AJ Drummond and A Rambaut), which includes the algorithm described in this paper, is available from http://evolve.zoo.ox.ac.uk/beast, with full installation and requirement details. The data set used in this paper is avaliable (see Additional file: 1)

Authors' contributions

IM conjectured and GL proved the one-state recursion. GL and IM independently implemented the algorithms, and wrote the paper. JLJ simplified the proof of the recursion, GL suggested to use it within an MCMC phylogeny cosampler, and IM suggested to use a Metropolised importance sampler and proved its correctness. GL and AD interfaced the Java algorithms to the BEAST phylogeny sampling package [40], and AD carried out the MCMC analysis. JH provided project management. All authors read and approved the final manuscript.

Additional material

Additional File 1

This XML file specifies the MCMC run for the example phylogeny and alignment co-estimation given in this paper (see Figs. 4, 5, 6). To run, download the BEAST package (AJ Drummond and A Rambaut, http://evolve.zoo.ox.ac.uk/beast.)

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-83-S1.xml]

Acknowledgements

The authors thank Yun Song, Dirk Metzler, Anton Wakolbinger and Ian Holmes for several useful suggestions and discussions. This research is supported by EPSRC (code HAMJW) and MRC (code HAMKA). I.M. was further supported by a Békésy György postdoctoral fellowship.

References

1. Thompson J, Higgins D, Gibson T: **CLUSTAL-W: improving the sensitivity of multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choise.** *Nucleic Acids Res* 1994, **22:**4673-4680.
2. Notredame C, Higgins D, Heringa J: **T-Coffee: A novel method for multiple sequence alignments.** *Journal of Molecular Biology* 2000, **302:**205-217.
3. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
4. Swofford D: **PAUP* 4.0.** *Sinauer Associates* 2001.
5. Felsenstein J: **PHYLIP version 3.63.** *Dept of Genetics, Univ of Washington, Seattle* 2004.
6. Sankoff D, Morel C, J CR: **Evolution of 5S RNA and the non-randomness of base replacement.** *Nature New Biology* 1973, **245:**232-234.
7. Jukes TH, Cantor CR: **Evolution of Protein Molecules.** In *Mammalian Protein Metabolism* Edited by: Munro. Acad Press; 1969:21-132.
8. Whelan S, Lió P, Goldman N: **Molecular phylogenetics: state-of-the-art methods for looking into the past.** *Trends in Gen* 2001, **17:**262-272.
9. Bishop M, Thompson E: **Maximum likelihood alignment of DNA sequences.** *J Mol Biol* 1986, **190:**159-165.

10. Thorne JL, Kishino H, Felsenstein J: **An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences.** *J Mol Evol* 1991, **33:**114-124.
11. Steel M, Hein J: **Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree.** *Appl Math Let* 2001, **14:**679-684.
12. Hein J: **An algorithm for statistical alignment of sequences related by a binary tree.** *Pac Symp Biocomp, World Scientific* 2001:179-190.
13. Holmes I, Bruno WJ: **Evolutionary HMMs: a Bayesian approach to multiple alignment.** *Bioinformatics* 2001, **17(9):**803-820.
14. Hein J, Jensen JL, Pedersen CNS: **Recursions for statistical multiple alignment.** *PNAS* 2003, **100(25):**14960-14965.
15. Miklós I: **An Improved Algorithm for Statistical Alignment of Sequences related by a Star Tree.** *Bul Math Biol* 2002, **64:**771-779.
16. Lunter G, Miklós I, Song Y, Hein J: **An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees.** *J Comp Biol* 2003, **10(6):**869-889.
17. Miklós I, Lunter GA, Holmes I: **A "Long Indel" model for evolutionary sequence alignment.** *Mol Biol Evol* 2004, **21(3):**529-540.
18. Holmes I: **A probabilistic model for the evolution of RNA structure.** *BMC Bioinf* 2004, **5(166):**.
19. Kuhner MK, Yamato J, Felsenstein J: **Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling.** *Genetics* 1995, **140(4):**1421-1430.
20. Griffiths RC, Tavare S: **Ancestral inference in population genetics.** *Stat Sci* 1994, **9:**307-319.
21. Wilson IJ, Balding DJ: **Genealogical Inference From Microsatellite Data.** *Genetics* 1998, **150:**499-450.
22. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W: **Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data.** *Genetics* 2002, **161(3):**1307-1320.
23. Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A: **The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach.** *Mol Biol Evol* 2003, **20(3):**381-387.
24. Felsenstein J: **Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregating sites as compared to phylogenetic estimates.** *Genetical Research Cambridge* 1992, **59:**139-147.
25. Stephens M, Donnelly P: **Inference in Molecular Population Genetics.** *J of the Royal Stat Soc B* 2000, **62:**605-655.
26. Pybus OG, Rambaut A, Harvey PH: **An integrated framework for the inference of viral population history from reconstructed genealogies.** *Genetics* 2000, **155(3):**1429-1437.
27. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
28. Jensen J, Hein J: **Gibbs sampler for statistical multiple alignment.** *Tech Rep 429, Dept of Theor Stat, U Aarhus* 2002.
29. Metzler D, Fleißner R, Wakolbringer A, von Haeseler A: **Assessing variability by joint sampling of alignments and mutation rates.** *J Mol Evol* 2001, **53:**660-669.
30. Metzler D: **Statistical alignment based on fragment insertion and deletion models.** *Bioinformatics* 2003, **19(4):**490-499.
31. Lunter G, Miklós I, Drummond A, Jensen J, Hein J: **Bayesian phylogenetic inference under a statistical indel model.** *Lecture Notes in Bioinformatics* 2003, **2812:**228-244.
32. Casella G, Robert CP: **Rao-Blackwellisation of sampling schemes.** *Biometrika* 1996, **83:**81-94.
33. Hein J, Wiuf C, Knudsen B, Møller MB, Wibling G: **Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit.** *J Mol Biol* 2000, **302:**265-279.
34. Miklós I, Toroczkai Z: **An improved model for statistical alignment.** *Lecture Notes on Computer Science* 2001, **2149:**1-10.
35. Dress A, Morgenstern B, Stoye J: **The number of standard and of effective multiple alignments.** *App Math Lett* 1998, **11(4):**43-49.
36. Liu JS: *Monte Carlo Strategies in Scientific Computing* Springer; 2001.
37. Giegerich R, Meyer C, Steffen P: **A Discipline of Dynamic Programming over Sequence Data.** *Science of Computer Programming* 2004, **51(3):**215-263.
38. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis* Cambridge University Press; 1998.
39. Holmes I, Durbin R: **Dynamic programming alignment accuracy.** *J Comp Biol* 1998, **5:**493-504.
40. Drummond AJ, Rambaut A: **BEAST v1.2.2.** 2004 [http://evolve.zoo.ox.ac.uk/beast].
41. Hedges SB, Poling LL: **A molecular phylogeny of reptiles.** *Science* 1999, **283(5404):**945-946.