

StatAlign: An Extendable Software Package for Joint Bayesian Estimation of Alignments and Evolutionary Trees

Ádám Novák,¹ István Miklós^{1,*}, Rune Lyngsø,¹ and Jotun Hein¹

¹Department of Statistics, University of Oxford, 1 South Parks Road, OX1 3TG Oxford, UK

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: Bayesian analysis is one of the most popular methods in phylogenetic inference. The most commonly used methods fix a single multiple alignment and consider only substitutions as phylogenetically informative mutations, though alignments and phylogenies should be inferred jointly as insertions and deletions also carry informative signals. Methods addressing these issues have been developed only recently and there has not been so far a user-friendly program with a graphical interface that implements these methods.

Results: We have developed an extendable software package in the Java programming language that samples from the joint posterior distribution of phylogenies, alignments and evolutionary parameters by applying the Markov chain Monte Carlo method. The package also offers tools for efficient on-the-fly summarisation of the results. It has a graphical interface to configure, start and supervise the analysis, to track the status of the Markov chain and to save the results. The background model for insertions and deletions can be combined with any substitution model. It is easy to add new substitution models to the software package as plugins. The samples from the Markov chain can be summarised in several ways, and new postprocessing plugins may also be installed.

Availability: The code is available from <http://phylogeny-cafe.elte.hu/StatAlign/>

Contact: miklosi@ramet.elte.hu

1 INTRODUCTION

The fundamental types of mutations that change biological sequences are substitution, insertion and deletion. Although insertions and deletions play an important role in the evolution of sequences, phylogenetic inference is often carried out taking only substitutions into account. Furthermore, analyses based on a single multiple alignment can be misleading, and multiple alignment methods tend to bring more variation to the phylogenetic analysis than tree building methods (Goldman, 1998; Wong *et al.*, 2008). Therefore, it is desirable to incorporate insertions and deletions in the phylogenetic analysis and to co-estimate phylogeny and alignment from their joint posterior distribution.

Time-continuous Markov models have been the standard for modelling substitutions in biological sequences (Jukes & Cantor,

1969; Felsenstein, 1981; Whelan *et al.*, 2001), and most of the popular phylogenetic inferring methods are based on such models (Ronquist & Huelsenbeck, 2003). Time-continuous Markov models for insertions and deletions have been developed by Thorne *et al.* (1991) and Thorne *et al.* (1992) that can also be used in phylogenetic analysis. Such analyses can highlight homoplasy and alignment uncertainty (Lunter *et al.*, 2005; Redelings & Suchard, 2005); can be applied for protein structure prediction (Miklós *et al.*, 2008) or phylogeny estimation of rapidly emerging pathogens (Redelings & Suchard, 2007).

Software packages published so far that fulfil some of the above described purposes (Holmes & Bruno, 2001; Fleißner *et al.*, 2005; Suchard & Redelings, 2006) lack a graphical interface and the potential of easy extension by further model and data summarisation plugins. We have implemented a package in the Java programming language that both has an easy-to-use user interface and is dynamically extendable through postprocessing and substitution model plugins – without any need for recompilation.

2 THE STATALIGN SOFTWARE PACKAGE

The main features of the program

The StatAlign software package is for joint Bayesian analysis of multiple alignments, phylogenetic trees and evolutionary parameters. The background model for insertions and deletions is a modified version of the TKF92 model (Thorne *et al.*, 1992) as described in Miklós *et al.* (2008). The indel model can be coupled with an arbitrary substitution model. We provide a wide selection of substitution models both for protein and nucleotide sequence data ranging from the Jukes-Cantor model to the general reversible nucleotide substitution model, and from the Dayhoff model to the WAG model (see the Model menu and/or the documentation).

The Bayesian analysis is based on Markov chain Monte Carlo (MCMC) employing the transition kernels described in Miklós *et al.* (2008). We made a few algorithmic improvements in the transition kernel for proposing tree topologies which effectively speeds up the analysis by a factor of 3-5. The new method proposes better alignments in a faster way, so it increases the acceptance ratio as well as decreases the amount of time needed for an MCMC step.

StatAlign has a graphical interface for choosing the input sequences, selecting the preferred substitution model and input-output formats, and setting MCMC parameters. During the analysis users can follow the progress through tabulated panels showing the

*to whom correspondence should be addressed

log-likelihood trace of the Markov chain to verify its convergence, the multiple alignment and phylogeny represented by the current state of the Markov chain, and the current Maximum Posterior Decoding estimate for the consensus alignment based on the sampled multiple alignments (see Fig. 1).

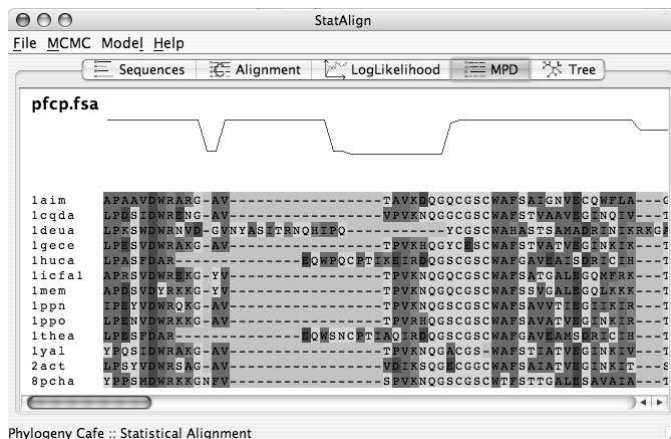


Fig. 1. A screenshot showing part of the current estimate of Maximum Posterior Decoding alignment during an MCMC run. 13 sequences are analysed that belong to the Papain family cysteine proteinases. Characters are coloured based on their chemical properties. The posterior decoding values are plotted above the alignment.

Modularity of the software package

Our aim was to build a software package with a fixed insertion-deletion model that can be coupled with an arbitrary substitution model. Due to its modularity, it is very easy to develop additional substitution models. We provide detailed description for developers to implement and plug-in new substitution models.

StatAlign generates random samples from the joint posterior distribution of sequence alignments, evolutionary trees and model parameters. This high-dimensional joint distribution can be analysed in several ways: the possibilities range from the simple statistics of marginalised single dimensions (e.g. the posterior distribution of a single rate parameter) to the covariation analysis of multiple dimensions. Besides, the convergence of the Markov chain might also be subject to investigations which vary from plotting of the log-likelihood trace to sophisticated analysis of autocorrelations. We provide detailed descriptions for developers to implement further postprocessing modules that perform such analyses and to visualise the results.

3 DISCUSSION

In this paper, we introduced a new software package with a graphical interface for the joint Bayesian estimation of alignments, phylogenies and evolutionary parameters. With a new transition kernel for proposing tree topologies, our program is significantly faster than the previously published version (Miklós *et al.*, 2008): one million MCMC steps on 13 sequences of papain family cysteine proteinases with an average length of 223 can be performed in less than 20 hours. Furthermore, the convergence of the Markov chain

is relatively fast: based on the loglikelihood trace, 10000 steps were sufficient for convergence on this dataset. Another novel feature of the program is that it is easily extendable: new substitution models as well as postprocessing modules can be plugged into the package without recompilation.

Bayesian phylogenetic inference is one of the most popular methods for analysing biological sequences. The standard protocol so far has been to align sequences with an alignment tool such as Clustal-W or T-COFFEE, and then use the alignment as the input for a program that considers only substitutions, e.g. MrBayes (Ronquist & Huelsenbeck, 2003). In contrast, our program allows joint inference of alignment, phylogeny and model parameters. This eliminates some artefacts that previous protocols suffer from, for example, that the tree estimated from the sequence alignment is influenced by the guide-tree that the alignment-building program used.

FUNDING

This research was supported by BBSRC grant BB/C509566/1. IM was also supported by a Bolyai postdoctoral fellowship and an OTKA grant F61730.

REFERENCES

- Bradley, R.K., Holmes, I. (2007) Transducers: An Emerging Probabilistic Framework for Modeling Indels on Trees, *Bioinformatics*, **23**, 3258–3262.
- Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998) Biological sequence analysis. Probabilistic models of proteins and nucleic acids, *Cambridge University Press*
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.*, **17**, 368–376.
- Fleißner, R., Metzler, D., von Haeseler, A. (2005) Simultaneous Statistical Multiple Alignment and Phylogeny Reconstruction. *Syst. Biol.*, **54**, 548–561.
- Goldman, N. (1998) Phylogenetic information and experimental design in molecular systematics. *Proceedings of the Royal Society London B*, **265**, 1779–1786.
- Holmes, I. (2003) Using guide trees to construct multiple-sequence evolutionary HMMs, *Bioinformatics*, **19**, i147–i157.
- Holmes, I., Bruno, W.J. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment, *Bioinformatics*, **17**, 803–820.
- Holmes, I., Durbin, R. (1998) Dynamic programming alignment accuracy, *J. Comp. Biol.*, **5**, 493–504.
- Lunter, G., Miklós, I., Drummond, A., Jensen, J.L., Hein, J. (2005) Bayesian coestimation of phylogeny and sequence alignment, *BMC Bioinformatics*, **6**, 83.
- Jukes, T.H., Cantor, C.R. (1969) Evolution of protein molecules, in *Mammalian protein metabolism* (ed.: H. N. Munro), **21–123**, Academic Press, New York.
- Miklós, I., Lunter, G.A., Holmes, I. (2004) A 'long indel' model for evolutionary sequence alignment *Mol. Biol. Evol.*, **21**, 529–540.
- Miklós, I., Novák, Á., Dombai, B., Hein, J. (2008) How reliably can we predict the reliability of protein structure predictions? *BMC Bioinformatics*, **9**, 137.
- Redelings, B.D., Suchard, M.A. (2005) Joint Bayesian Estimation of Alignment and Phylogeny. *Systematic Biology*, **54**(3):401–418.
- Redelings, B.D., Suchard, M.A. (2007) Incorporating indel information into phylogeny estimation for rapidly emerging pathogens, *BMC Evolutionary Biology*, **7**, 40.
- Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models *Bioinformatics*, **19**, 1572–1574.
- Suchard, M.A., Redelings, B.D. (2006) BALI-Phy: simultaneous Bayesian inference of alignment and phylogeny, *Bioinformatics*, **22**, 2047–2048.
- Thorne, J.L., Kishino, H., Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences *J. Mol. Evol.*, **33**, 114–124.
- Thorne, J.L., Kishino, H., Felsenstein, J. (1992) Inching toward reality: an improved likelihood model of sequence evolution, *J. Mol. Evol.*, **34**, 3–16.
- Whelan, S., Liò P. and Goldman, N. (2001) Molecular phylogenetics: state of the art methods for looking into the past. *Trends in Genetics* **17**:262–272.
- Wong, K.M., Suchard, M.A., Huelsenbeck, J.P. (2008) Alignment uncertainty and genomic analysis, *Science*, **319**, 473–476.