

Improved Bayesian Phylogenetic Inference in a Statistical Alignment Framework

Advanced Software Design for StatAlign

Ádám Novák, István Miklós, Rune Lyngsø, Jotun Hein

24 February 2011

1 Introduction

Long-term trends in computational phylogenetics show a steady transition of focus from traditional tree reconstruction methods towards Bayesian approaches. The early distance based techniques such as UPGMA and Neighbour Joining are today considered less accurate primarily due to the loss of information when condensing sequence data into a distance matrix. Maximum parsimony is fast but suffers from a number of issues: it is not statistically consistent, does not consider all possible evolutionary histories and is prone to the artifact of long branch attraction. Maximum likelihood methods on the other hand have the power to fully capture all possible evolutionary scenarios but can be prohibitively slow. Tools implementing this and before mentioned strategies include PAUP [16], PAML [20] and PHYLIP [3]. With the advent of Bayesian phylogenetic methods analysing a moderately large set of sequences in one go has become possible while retaining solid statistical grounds. A standard tool in this area is MrBayes [7, 14]. For reviews on the subject see e.g. [5] and [6].

Despite their wide usage, all of the classic methodologies discussed so far share an intrinsic weakness: the alignment and phylogenetic tree reconstruction phases are separated and the latter is based on a single, fixed alignment. This prevents alignment ambiguity from being taken into account when constructing the tree [11, 19] while the alignment algorithm must do without a good estimate of the evolutionary relationships of the sequences. Also, it has been shown that the output of most phylogeny reconstruction algorithms are biased towards the guide tree that was used to construct the alignments when sequences were aligned using a progressive approach [17]. These observations led to the development of strategies for joint Bayesian estimations of alignment and phylogeny, including BAli-Phy [15] and StatAlign [13].

These methods operate in a Bayesian framework and employ MCMC to sample from the joint posterior space of alignments, trees and model parameters. They both work with a realistic insertion–deletion model on the gene

level, TKF92 [18] and offer a selection of substitution models. The clear advantage of this approach is that the complete evolutionary history is modelled and estimated in a single, sound statistical framework.

2 Project description

Our recently created Statistical Alignment package, StatAlign [13] is written in Java and was designed to be easily extendable with postprocessing plugins to analyse posterior samples. The most recent version (v1.1) offers powerful tools to summarise alignment samples in a representative alignment and although also provides samples of evolutionary trees, it lacks the necessary functionality to aid advanced phylogenetic studies. Besides, the current single-threaded implementation does not support the utilisation of multiple processors to speed up the sampling process. We expect that the increased accuracy that is offered by the underlying evolutionary model can be directly translated into more rigorous phylogenetic inferences while making use of information inherent in sequence data to the extremes. Therefore, this project is looking to equip StatAlign with missing features to assist phylogenetic analyses and thus help promote it as a cutting edge phylogenetic package.

2.1 Consensus trees and networks

There are well established algorithms to construct a consensus tree given posterior samples of the tree space. A *consensus tree* is a single representative tree that includes features that all or most of the trees agree on. Each sampled tree is a statement about the relationships of multiple groups of taxa, more specifically, each branch of each tree splits the set of all taxa into two partitions, both of which contain taxa that are more closely related to each other than to taxa in the other partition. The set pairs emerging this way are referred to as *bipartitions* (see Figure 1).

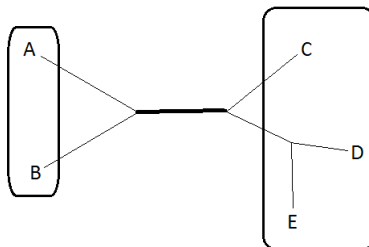


Figure 1: The bipartition defined by the thickened edge in an unrooted tree. The topology of the two subtrees is irrelevant.

A *strict consensus tree* is a tree that contains all bipartitions that have been observed in all samples and leaves the rest of the branches unresolved. This is

a bit too crude and sometimes results in a tree with very few internal nodes. The *majority rule consensus tree* contains all bipartitions that are present in the majority (larger than 50%) of the trees. These bipartitions are never conflicting: any bipartition that is incompatible with a majority one naturally cannot appear in a tree where the majority bipartition is present (which is more than 50% of all trees) and thus the incompatible one cannot be a majority bipartition. This proves that the majority rule tree always exists. Some implementations leave the remaining branches unresolved or it is common to consider the remaining bipartitions in a decreasing order of their frequency, only adding the ones that are compatible with the ones fixed earlier. Figure 2 shows an example when the majority rule tree is fully resolved.

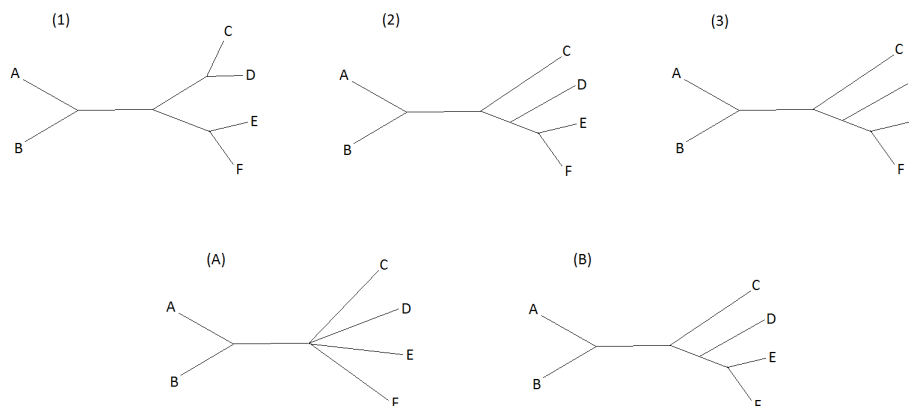


Figure 2: Three sampled trees (1,2,3) and the corresponding *strict* (A) and *majority rule* (B) consensus tree. The bipartition $AB|CDEF$ is present in all samples and apart from that, $ABCD|EF$ and $ABC|DEF$ are in the majority (two) of the trees.

In accordance with StatAlign’s philosophy of allowing users to monitor the progress of the sampling process it is the goal of the present project to develop a plugin that periodically calculates and visualises estimates of the consensus tree based on a subset of samples. This must not have a significant impact on the total running time so the algorithms summarising samples must be very efficient. A good starting point is the $O(tn)$ algorithm by Amenta et al. [2] to compute the majority rule tree where t is the number of taxa and n is the number of trees.

A tree is not the only conceivable structure that could capture the posterior space of phylogenies in a representative way. In particular, consensus networks have proven to be useful in marking the ambiguous regions of the trees and are more informative [9]. Efficient algorithms have been designed which condense a set of trees into a network [10] that are able to illustrate even otherwise incompatible splits (in the phylogenetic network context bipartitions are typically

referred to as splits). An example is shown in Figure 3. Most of these methods are implemented in SplitsTree [8].

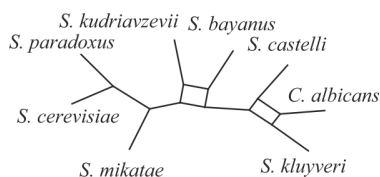


Figure 3: An example splits network from Huson et al. [10] representing all splits that occur in more than 30% of the original 106 trees on eight yeast species.

The plugin to be developed shall support the calculation of both consensus trees and networks and shall include powerful visualisation tools. Though this is mainly a software project it is also necessary to test the new features on sequence data where it is expected that alignment uncertainty will have significant effect on the topology/reliability of the predicted phylogeny, to demonstrate how the statistical alignment approach can be superior in these scenarios. This will require relevant data (to be downloaded from sequence databases) and running existing phylogenetic programs.

A short project could concentrate on development suggestions up to this point that could be undertaken in a 6-week period.

2.2 Usability improvements

The proposed extensions in the following sections will be the target of extended projects. Phylogenetic studies are often based on data coming from different sources: protein sequences, nucleotide sequences from mitochondrial, plastid or nuclear genomes or even morphological characters. Support for composite data at StatAlign’s model level would be beneficial in multiple ways: it would allow phylogenetic inference from all available data in one step and give statistically consistent estimates of the consensus tree (rather than forcing the user to come up with ad-hoc methods to combine results from separate analyses).

This improvement calls for a few major changes to the StatAlign core and the development of evolutionary models for new data types, especially morphological characters. The application GUI shall be reorganised to reflect the new structure and aid easy setup of running parameters. MrBayes supports mixed models starting from version 3 [14] and is a good source of inspiration.

The main development phase of this subproject could be fitted into another 6-week period while assessment of the mixed model inference and comparison to existing tools requires an additional 2-3 weeks.

2.3 Parallelisation

Markov chain Monte Carlo techniques have proven to be powerful when direct sampling from a posterior target distribution is not feasible. Although the Metropolis–Hastings algorithm guarantees that the global optimum will eventually be found [12] it may perform poorly in multimodal scenarios when high-probability peaks are separated by large low-probability regions, in which case jumps between the peaks occur too infrequently and mixing becomes inefficient.

Multi-chain MCMC methods such as Metropolis coupled MCMC (or (MC)³) have been proposed to address the issue [4]. The essential underlying idea is that several Markov chains are run in parallel, most of which are 'heated' to some degree, meaning that the probability distribution is flattened, approaching a uniform distribution in the limit. This can be achieved by raising the posterior probability density function $f(P | X)$ to the power β : a heated chain will explore the distribution $f(P | X)^\beta$ where $0 < \beta < 1$ is the heat parameter ($\beta = 1$ for the cold chain). These heated chains allow rapid jumps between very distant peaks, thus eliminating the slowdown effect of low-probability valleys. Occasionally swaps are attempted between chains at different levels, which will effectively speed up mixing of the cold chain, too.

While traditional MCMC algorithms are difficult to parallelise due to the intrinsic property of Markov chains that each state depends on the previous (few) states, Metropolis coupled MCMC can be very efficiently implemented on parallel architectures. The chains can be run separately for long periods and synchronisation of the processors is only necessary when swaps are done between the chains. Efficient parallel (MC)³ variants have been proposed for phylogenetic inference by Altekar et al. [1] and are implemented in both MrBayes [14] and BAli-Phy [15].

This extended subproject is aiming to add (MC)³ support to StatAlign to boost mixing of Markov chain when large datasets are being analysed and to make best use of multi-core CPUs that are standard in today's computers.

3 Skills required

The project can be carried out by a student or a group of students with good Java programming experience and solid algorithmic and maths background. Biological knowledge is not crucial though a general overview of molecular biology is useful.

References

- [1] Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., Ronquist, F. (2003). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, doi:10.1093/bioinformatics/btg427.

- [2] Amenta, Nina, Clarke, Frederick, St. John, Katherine (2003). A randomized linear-time majority tree algorithm. In *Proceedings of the European Conference on Computational Biology*, (2003).
- [3] Felsenstein, Joseph (1989). PHYLIP - Phylogeny Inference Package (version 3.2). *Cladistics*, **5**:164–166.
- [4] Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface* (Ed. Keramidas, E. M.), 156–163., Interface Foundation.
- [5] Holder, Mark, Lewis, Paul O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, **4**(4):275–284.
- [6] Huelsenbeck, J., Ronquist, F., Nielsen, R., Bollback, J. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**:2310–2314.
- [7] Huelsenbeck, J. P., Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**(8):754–755.
- [8] Huson, D H (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**(1):68–73.
- [9] Huson, D. H., Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**(2):254–267.
- [10] Huson, Daniel H., Scornavacca, Celine (2011). A survey of combinatorial methods for phylogenetic networks. *Genome Biology and Evolution*, **3**:23–35.
- [11] Lutzoni, F., Wagner, P., Reeb, V., Zoller, S. (2000). Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Systematic Biology*, **49**(4):628–651.
- [12] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6):1087–1092.
- [13] Novák, Á., Miklós, I., Lyngsø, R., Hein, J. (2008). StatAlign: An extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, **24**(20):2403–2404.
- [14] Ronquist, F., Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**(12):1572–1574.
- [15] Suchard, M. A., Redelings, B. D. (2006). BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**(16):2047–2048.
- [16] Swafford, D. (2002). PAUP*. Phylogenetic analysis using parsimony (*and other methods). version 4.

- [17] Thorne, J. L., Kishino, H. (1992). Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution*, **9**(6):1148–1162.
- [18] Thorne, J. L., Kishino, H., Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, **34**:3–16.
- [19] Wong, K. M., Suchard, M. A., Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, **319**(5862):473–476.
- [20] Yang, Z. (1997). PAML: A program package for phylogenetic analysis by Maximum Likelihood. *Computer applications in the biosciences*, **13**(5):555–556.