

Assessing the convergence properties of MCMC samplers for statistical alignment

J. L. Herman, Á. Novák and J. Hein*

Department of Statistics, University of Oxford, United Kingdom

Multiple sequence alignment has long been one of the key problems in bioinformatics, and is an essential prerequisite for a whole range of downstream analyses, ranging from phylogenetic inference, to homology modelling. Statistical models for sequence evolution with insertions and deletions were first developed at the beginning of the 90s, and enabled alignment to be performed within a probabilistic framework [1]. For certain models, it is possible to use dynamic programming algorithms to efficiently compute the sum over alignments [2], allowing the maximum likelihood (or *maximum a posteriori*) alignment to be computed, as well as enabling alignments to be sampled directly from the posterior. However, when the number of sequences increases, this analytic treatment rapidly becomes intractable, and it is necessary to resort to methods such as Markov chain Monte Carlo (MCMC) to sample this space [3].

While the MCMC approach enables us to explore the posterior distribution over alignments without computing the normalising constant, the space of alignments is so vast that on average each observed alignment is never sampled more than once, such that convergence to the stationary distribution over alignments is exceedingly unlikely. In order to tackle this problem, we typically focus on the marginal frequencies for each observed alignment *column*, summed over all alignments. Because the space of columns is many orders of magnitude smaller than the space of alignments, it is possible to estimate these marginals more reliably [4].

Nevertheless, in order to assess the reliability of the samples we obtain from a particular statistical alignment run, it is desirable to more formally assess both the rate of convergence to the stationary distribution, and the fraction of the posterior mass explored by the MCMC sampler. This project will seek to conduct investigations into such properties, for which purpose it is necessary to develop effective summary statistics to describe the high dimensional alignment space. By analysing a test case where we are still able to compute the exact posterior analytically, it will be possible to examine in detail the performance of the MCMC-based methods, which will ultimately allow for an automated method for assessing the convergence of a particular MCMC run.

References

- [1] Thorne, J. L., Kishino, H. and Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, **33**(2):114–124
- [2] Lunter, G. A., Miklós, I., Song, Y. S. and Hein, J. (2003) *Journal of Computational Biology*, **10**(6):869–889
- [3] Lunter, G. A., Miklós, I., Drummond, A., Jensen, J. L. and Hein, J. (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**:83
- [4] Herman, J. L., Novák, A., Miklós, I., Lyngsø, R., and Hein, J. (*submitted*). Efficient posterior decoding for statistical alignments.

*hein@stats.ox.ac.uk