Efficient calculation of a representative multiple sequence alignment from posterior samples

Joseph L. Herman*, Adám Novák, Rune Lyngsø and Jotun Hein

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

A number of methods have recently emerged for probabilistic alignment of protein sequences under a Bayesian framework. A common feature of many such methods is the use of Markov chain Monte Carlo techniques to explore the posterior space of alignments. However, no clear consensus exists as to how to interpret the output of such simulations in such a way as to facilitate comparison with traditional procedures that yield a single alignment, particularly when multiple sequences are involved. Due to the size of the alignment space, each observed alignment is typically sampled with a very low frequency, such that the maximum a posteriori alignment is often a poor summary of the posterior.

Working instead in the space of alignment columns, each column may be observed within a large number of different alignments, such that the marginal posterior for each observed column can be estimated much more reliably from a simulation. We develop a framework for coding the space of sampled alignments in order to permit a directed acyclic graph representation of the posterior in the space of columns; each path through the graph then represents a valid alignment. This representation allows the computation of globally consistent marginal posterior probabilities for the presence of each column in the alignment given observed frequencies, and enables the use of shortest-path algorithms for calculating single alignments that minimise the posterior risk under a particular family of loss functions. This allows us to generate representative alignments that reflect a desire to minimise false positives or false negatives to varying degrees, depending on prior information regarding the expected homology between the sequences. Using a length-independent measure of alignment accuracy, we compare the results of the methodology with manually curated alignments for a selection of protein families.

*herman@stats.ox.ac.uk