STEME: Efficient EM for motif search using suffix trees

March 3, 2011

John Reid - MRC Biostatistics Unit, Cambridge, UK - john.reid@mrc-bsu.cam.ac.uk

Lorenz Wernisch - MRC Biostatistics Unit, Cambridge, UK

MEME and many other popular motif finders use the expectation-maximisation (EM) algorithm to optimise their parameters. Unfortunately the running time of EM is linear in the length of the input sequences. This can prohibit its application to data sets of the size commonly generated by high-throughput biological techniques. A suffix tree is a data structure that can efficiently index a set of sequences. We describe an algorithm, Suffix Tree EM for Motif Elicitation (STEME), that approximates EM using suffix trees. To the best of our knowledge this is the first application of suffix trees to EM. We provide an analysis of the expected running time of the algorithm and demonstrate that STEME runs an order of magnitude more quickly than the implementation of EM used by MEME. We give theoretical bounds for the quality of the approximation and show that, in practice, the approximation has a negligible effect on the outcome. We provide an open source implementation of the algorithm that we hope will be used to speed up existing and future motif search algorithms at http://sysbio.mrc-bsu.cam.ac.uk/johns/STEME/.