

# Applications of the EM Algorithm in Molecular Biology

## Abstract

In this paper, I will provide a detailed explanation on the EM Algorithm, along with a comprehensive description about the Jensen's Inequality; which is a key theory used in the algorithm. Furthermore, I will discuss 2 popular applications of the EM Algorithm in Life Science: Motif Finding, and Linkage Analysis. The method in which the EM Algorithm is applied to these 2 areas will be covered in detail as well. Following that, I will introduce an extension of the EM Algorithm for Motif Finding - MEME. Lastly, I will cover the application of the EM Algorithm and MEME, with respect to Motif Finding in Herpes Viruses DNA. The results and Findings will be discussed.

While writing this paper, I had in mind to structure the content for undergraduates in Statistics, and for any other readers who may be interested. Hence, the explanations of certain theories and concepts will not assume that the reader has background knowledge, and thus be described step-by-step, supported with figures when possible.

The EM Algorithm R code was written by me, and the code was specifically customized for the application to Motif Finding. It has been tested for reliability and validity by using multiple test sample sequences. The full code can be found in the Appendix. All figures were drawn by myself, with the exception of the MEME output in Chapter 5; cropped image of the results from the online MEME webpage. The detailed descriptions of the EM Algorithms in Chapter 3 and 4 were original as well. Finally, the results and conclusions of the findings were original too.

A0086098N

Yeun Jun Wei Samuel