An EM Algorithm for Binding Energy Estimation Using HT-SELEX Data

Shuxiang Ruan[1], S. Joshua Swamidass[2], Gary D. Stormo[1,*]

[1]*Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA.*
[2]*Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA.*

ABSTRACT
The interaction between transcription factors and DNA plays an important role in gene expression regulation. In this study, we developed an expectation-maximization (EM) algorithm, called EMSEL, for extracting binding motifs from high-throughput SELEX (HT-SELEX) data. EMSEL builds on a comprehensive biophysical model of protein-DNA interactions and is capable of estimating the confidence intervals of the parameters in the model. We compared the binding motifs generated by EMSEL with those estimated by other algorithms using both HT-SELEX and ChIP-seq data. The results demonstrate that the EMSEL motifs generate significantly better predictions of the *in vitro* data and their predictions of the *in vivo* data are comparable to the other motifs based on the criterion of the area under the ROC curve (AUC). The ChIP-seq test results, together with the fact that many of the non-EMSEL motifs have very high information content, highlight the limitations of the AUC criterion, which is purely rank-based and fails to take account of the relative binding affinities of ChIP-seq peaks.