# IMPROVED ENSPART FOR DNA MOTIF PREDICTION

[1]Allen Chieng Hoon Choong, [1,*]Nung Kion Lee, [2]Chih How Bong and [1]Norshafrina Omar
[1]Faculty of Cognitive Sciences and Human Development
[2]Faculty of Computer Science and Information Technology
University Malaysia Sarawak, 94300 Kota Samarahan
*Corresponding author: nklee@unimas.my

## ABSTRACT

In our previous work we proposed ENSPART-an ensemble method for DNA motif discovery which partitions input dataset into several equal size subsets runs by several distinct tools for candidate motif prediction. The candidate motifs obtained from different data subsets are merged to obtain the final motifs. Nevertheless, the original ENSPART has several limitations: (1) the same background sequences are used for the calculation of Receiver Operating Cost (ROC) of motifs obtained from different datasets. This causes bias because different datasets might have different background distribution; (2) it does not consider the duplication of a motif and its reverse complement. This causes many redundant motifs in the result set which requires filtering. In this work, we extended the original ENSPART to solve those two issues. For the first issue, we employed background sequences that is based on the distribution of bases in the input sequences. As for the second issue, we employ a "triple" merging strategy to reduce redundant motifs. Our evaluation results indicate that the two improvements obtain better AUC values in comparison to the original implementation.

## INTRODUCTION

ENSPART (Lee, Choong, & Omar, 2016) is an ensemble approach which utilizes an ensemble of 7 motif discovery tools for motif prediction. It is designed for tackling large-scale ChIP dataset for the discovery of primary motifs in a DNA dataset enriched with motifs. The idea of ENSPART is to partition a large-scale ChIP dataset into small subsets and use an ensemble of motif discovery tools for motif prediction in each subset. The assumption is the binding sites of a primary transcription factor protein is abundance in each of the partitioned subset and thus can be predicted by motif discovery tools independently. Furthermore, utilizing many tools for prediction would increase the chances of obtaining true motifs. The tools run on each partitioned dataset for motif discovery and predicted motifs from individual tool are merged to produce the final motifs. An alignment free method is employed to merge motifs obtained from different data subsets to reduce redundancy and groups similar motifs. The merging managed to reduce about 49 to 55% of the motifs produced for all the evaluated datasets. The receiver operating curve (ROC) is used to rank the candidate motifs before the final motifs selection. Our previous simulation results demonstrated ENSPART good performance in comparison to MEME. Nevertheless, the original implementation of ENSPART has several noticeable weaknesses:

- The calculation of ROC used for ranking of candidate motifs require a set of background sequences which does not contain the motifs. In our implementation, we employed the same background sequences for the computation of ROC for the ranking of final motifs from different datasets. This could be biased since it is not guaranteed the background sequences do not contain motifs.
- The existing merging method does not consider the similarity between the motifs in the forward and reverse complement. There could be many redundant motifs due to that.

In this paper, the improvements over the original ENSPART by addressing the two issues above will be presented.

This paper is organized as follows. The background section presents some background of DNA motif prediction problem and ENSPART algorithm. The next section presents the modifications