Random forest have been applied to predicting DNA binding sites of proteins and protein pairs in a new Master’s thesis

Anni Antikainen has developed methods for modeling protein-DNA binding sites in a new Master’s thesis titled Modeling protein-DNA binding specificities with random forest. The thesis has been written in the Department of Computer Science in the field of computational systems biology.

Random forest was found to predict DNA binding sites of single proteins with 71.6 % accuracy and protein pair binding sites with 79.0 % accuracy. Random forest is a method based on learning multiple decision trees on the training data. Furthermore, decision trees are models that partition the data according to local models repetitively. The random forest models were trained with DNA sequences known to comprise a binding site and random DNA sequences supposedly not including binding sites. Performance of the models were tested with unseen DNA sequences.

Proteins turn genes on and off by binding to specific DNA sequences near the target genes. Therefore, predicting DNA binding sites is crucial for understanding differences in gene expression between different conditions such as healthy and diseased individuals with the same genetic predisposition. Random forest predicted binding sites of individual proteins with 71.6 % accuracy, which is a significant improvement from the prediction accuracy of 63.3 % of the most popular binding preference models, position weight matrices. Furthermore, an other model called DeepBind have previously been proposed for predicting individual protein binding sites. Random forest performed almost equally to DeepBind, which predicted binding sites with 71.9 % accuracy. Thus, random forest models can reveal if DNA sequences comprise a binding site or not.

A random forest model was also implemented for predicting binding sites of protein pairs, since some proteins that regulate gene expression can bind different DNA sequences as pairs than they would have individually. Information about DNA binding sites of protein pairs is crucial for revealing gene expression patterns. Random forest predicted protein pair binding sites with 79.0 % accuracy when position weight matrices achieved 61.8 % accuracy. Since DeepBind is not provided for protein pair binding sites, random forest provide beneficial information about protein pair DNA binding preferences and can be utilized for predicting binding sites from DNA sequences.

Contact information: Anni Antikainen, anni.antikainen@aalto.fi