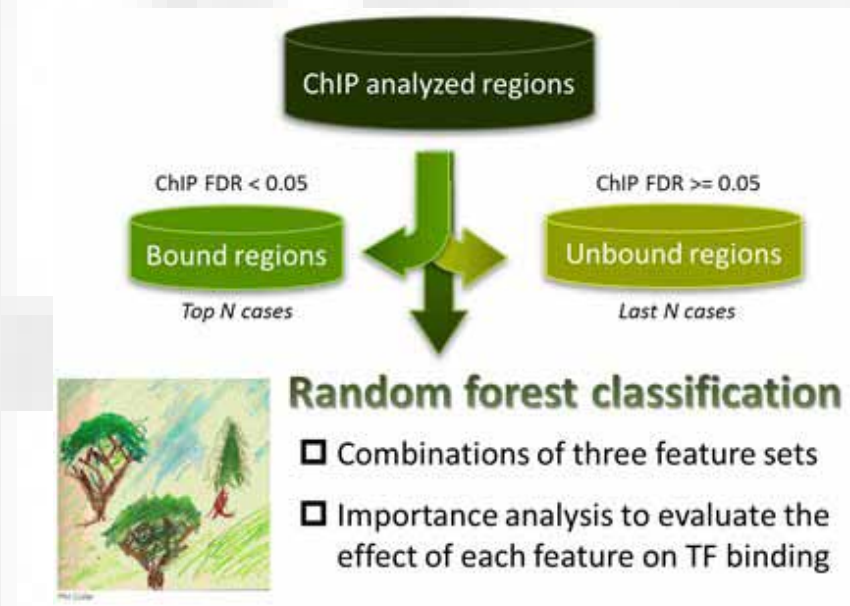
Bioinformatics Approach for Transcription Factor Binding Properties

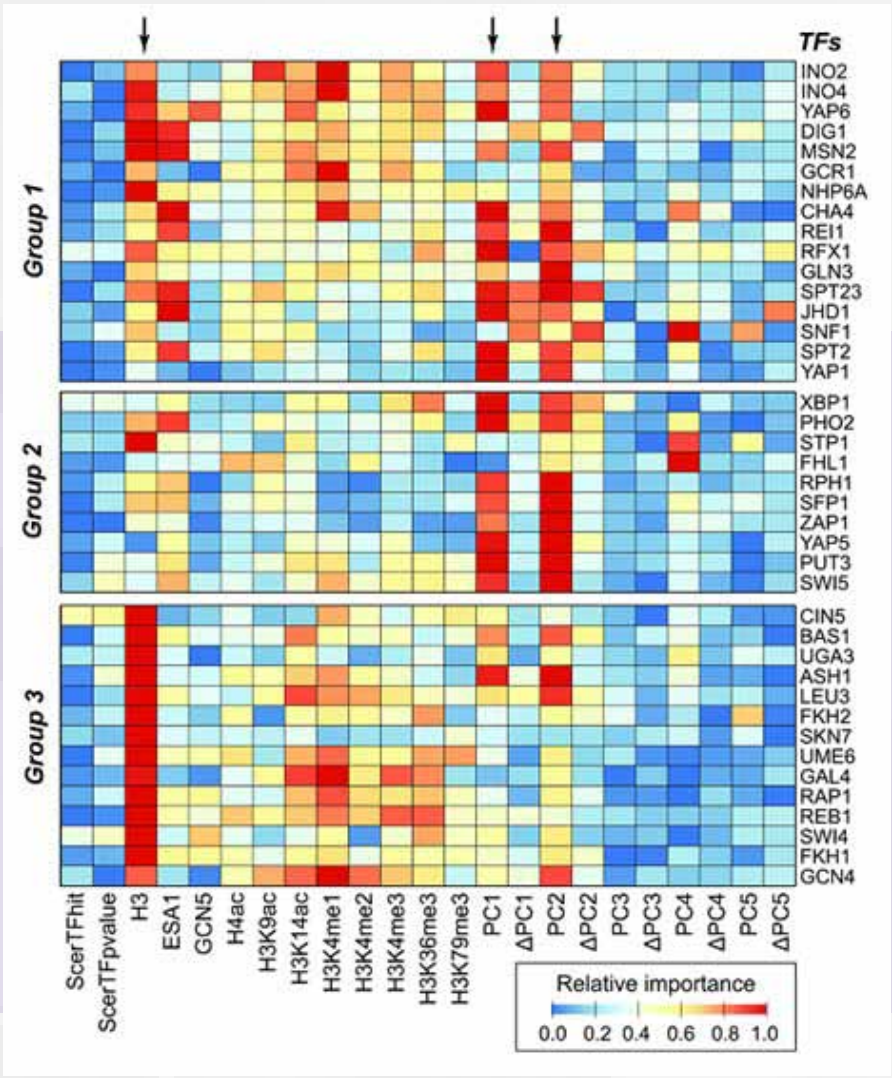
Huai-Kuang Tsai
Research Fellow



A random forest classifier for transcription factor binding properties. (Fig. 1).

One of the central questions in molecular genetics regards the mechanisms of transcriptional regulation, particularly how transcription factors (TFs) regulate expression of target genes with specific TF binding sites (TFBSs). Identifying TFBSs would permit a more comprehensive and quantitative mapping of the regulatory mechanisms within cells. Unfortunately, TFBSs are usually short (~5–20 bp) and degenerate, making it difficult to accurately identify TFBSs. With the advancement of biological techniques, there are widely applicable methods to identify TFBSs experimentally, such as Electrophoretic Mobility Shift Assay (EMSA) , Systematic Evolution of Ligands by Exponential Enrichment (SELEX), Chromatin Immunoprecipitation (ChIP) assays, and Protein Binding Microarrays (PBMs). Experimental methods provide in vivo evidence of TF binding or in vitro measurement of affinity of TF–DNA interactions. However, large-scale and precise prediction of TFBSs remains one of the greatest challenges due to high cost and low time efficiency. Genome-wide TFBS identification thus requires the complementation of bioinformatics.

A large number of studies have developed computational methods for TFBS prediction that examine the presence of sequence motifs, usually simplified as a motif-discovery problem, which involves seeking a sequence motif from a vast array of biological data, such as the promoter sequences of target genes. The motif is typically modeled as a position weight matrix (PWM). PWMs can be used to infer the binding strength of sequences based on the number of known TFBSs or potential binding sequences. Motif-discovery methods for TFBS identification usually follow an enumerative or probabilistic approach. Enumerative approaches investigate the occurrence frequency of all strings, and generate a PWM composed of over-represented strings.



The relative importance of three kinds of features, including sequence motif, chromatin state, and DNA structure, for predicting binding regions of different TFs. Arrowheads indicate the most important features for TFs. (Fig. 2).

Alternatively, probabilistic approaches conduct a multiple sequence alignment of input sequences and simultaneously optimize PWM parameters using machine learning methods, such as Expectation-Maximization algorithm and Gibbs sampling.

With the increase of available genome-wide data is greater evidence indicating that DNA sequence is not the only factor determining TF binding. In particular, studies have shown that a large portion of false positives in TFBS identification (i.e., motif occurrences which are not really bound by TFs) is due to chromatin inaccessibility. Furthermore, chromatin accessibility and TF binding affinity are found to be correlated. These observations reveal that chromatin accessibility could be a key factor that controls TF binding. Chromatin state feature and DNA structural property are the two main categories of genomic attributes associated with chromatin accessibility. Both chromatin state and DNA structural properties have been shown to be determinants of chromatin accessibility and consequently influence TF binding. Recently, certain TFBS identifications using chromatin state or DNA structural properties have been developed.

Although sequence motifs (SM), chromatin state (CS), and DNA structural (DS) properties have been used to predict TF binding sites, a predictive model that jointly considers CS and DS has not been developed to predict either TF-specific binding or general binding properties of TFs. Using budding yeast as model, we found that machine learning classifiers (random forest, as shown in Fig. 1) trained with either CS or DS features alone perform better in predicting TF-specific binding compared to SM-based classifiers. In addition, simultaneously considering CS and DS further improves the accuracy of the TF binding predictions, indicating the highly complementary nature of these two properties. The contributions of SM, CS, and DS features to binding site predictions differ greatly between TFs, allowing TF-specific predictions and potentially reflecting different TF binding mechanisms. In addition, a "TF-agnostic" predictive model based on three DNA "intrinsic properties" (in silico predicted nucleosome occupancy, major groove geometry, and dinucleotide free energy, see Fig. 2) that can be calculated from genomic sequences alone has performance that rivals the model incorporating experiment-derived data. This intrinsic property model allows prediction of binding regions not only across TFs but also across DNA-binding domain families with distinct structural folds. Furthermore, these predicted binding regions can help identify TF binding sites that have a significant impact on target gene expression. Because the intrinsic property model allows prediction of binding regions across DNA-binding domain families, it is TF agnostic and likely describes the general binding potential of TFs. Thus our findings suggest that it is feasible to establish a TF-agnostic model for identifying functional regulatory regions in potentially any sequenced genome.

Dr. Zing Tsung-Yeh Tsai, a postdoc at my lab; and Dr. Shin-Han Shiu, an associate professor of plant biology at Michigan State University, contributed to this work.