SEARCH OF CELL-SPECIFIC TRANSCRIPTION FACTOR BINDING SITES WITH DNASE HYPERSENSITIVITY AND HISTONE MODIFICATIONS

EG Gusmão[1], C Dieterich[2], IG Costa[1,3]

[1] IZKF & Institute for Biomedical Engineering, RWTH University Medical School, Germany.
[2] Max Delbrück Center for Molecular Medicine Berlin, Germany.
[3] Center of Informatics, Federal University of Pernambuco, Brazil.

Eukaryotic gene expression involves the coordination of a multitude of transcription factors that bind on specific cis-acting DNA elements. The understanding of the complex regulatory networks is crucial for the comprehension of biological processes such as cell differentiation and the onset of diseases. Standard sequence-based computational approaches to find binding sites suffers from a high number of false positive hits, given its incapability to identify active sites. Current research has proven that novel genome-wide assays reflecting chromatin structure, such as DNase I digestion (obtained with DNase-seq) or histone modifications (obtained with ChIP-seq) outperform sequence-based detection of transcription factor binding sites that are active in a particular cell type. Moreover, the discovery of distinct modes of chromatin signatures have strengthened these results and reinforced the usage of epigenetic datasets to such purpose.

Previous methods on detecting cell-specific binding sites from chromatin information can be categorized in two groups: site-centric and segmentation-based. Both use cell-specific experimental data but site-centric methods require sequence information to make factor-specific predictions while segmentation-based methods annotate the genome for likely binding locations for any transcription factor. In this study, multivariate hidden Markov models are used to detect transcription factor binding locations using genome-wide high-resolution DNase I digestion and histone modification data. In this segmentation-based methodology, a complete regulatory map can be generated with a single run using experimental data for a specific cell-type.

We used public data from ENCODE project to predict binding sites in the cell lines H1-hESC, HeLa-S3, HepG2 and K562. Moreover, we performed a comprehensive comparative analysis of competing methods using a well-established gold standard, which was generated by combining sequence motifs with ENCODE's ChIP-seq data for 22 factors. Our method outperformed all other competing methods concerning a Friedman-Nemenyi test applied to the default metric in this scenario: the area under the ROC curve. Furthermore, we observed that our strategy provides a better balance between sensitivity and specificity for detection of active binding sites. For instance, our method presented 83% sensitivity and 96% specificity regarding GABPA binding in H1-hESC, while the segmentation-based method by Boyle et al. achieved 59% sensitivity and 98% specificity. In contrast, site-centric methods were more sensitive but showed low specificity. For instance, Cuellar-Partida et al. achieved 87% sensitivity but only 84% specificity in the same previous scenario, the lowest among all tested methods. Additional experiments demonstrated that our method is independent of cell-type concerning the model's training and that it outperforms all other segmentation-based methods concerning their spatial accuracy, i.e. how close predicted regions are to actual binding sites. Finally, improvement of genome-wide high-resolution data quality as a result of advances in high-throughput sequencing techniques, gives rise to a number of potential applications to the framework presented here.

Supported by: Interdisciplinary Center for Clinical Research (IZKF), RWTH University Medical School, Aachen, Germany; and Brazilian research agencies: FACEPE and CNPq.