Predicting CTCF site occupancy using sequence and chromatin-associated features

Sunil Kumar[1,2], René Dreos[2], Giovanna Ambrosini[1,2], Philipp Bucher[1,2]
[1]*Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, EPFL, Station 15, Lausanne CH-1015, Switzerland*
[2]*Swiss Institute of Bioinformatics (SIB), EPFL, Station 15, Lausanne CH-1015, Switzerland*

Introduction

Genes are regulated by transcription factors (TF) binding to physiological target sites in the genome. Understanding the mechanisms by which TFs are recruited to their target sites is essential for the understanding of gene regulation. The recently introduced ChIP-Seq technology allows for genome-wide mapping of all *in vivo* bound sites of a given TFs in a particular cell type at near base-pair resolution [1]. What has become clear from ChIP-Seq experiments is that the intrinsic binding specificity of a TF can only partly explain the *in vivo* site occupancy patterns, which were found to be remarkably tissue-specific.

Several recent studies have reported that TF binding is influenced and can be chromatin contextual features such as DNA chromatin accessibility, nucleosome occupancy, or the presence of specific histone post-translational modifications [2,3]. Site occupancy may also be partly predictable from sequence intrinsic properties such as oligo-nucleotide composition, DNA structural parameters and evolutionary conservation. In this work, we use machine learning to assess the relative importance of such features in TF to target site recruitment process, in the hope to gain insights into transcription regulatory mechanisms. As test example, we use the sequence-specific DNA-binding protein CTCF which has been assayed by ChIP-Seq in many cell types. CTCF has been attributed diverse roles in gene regulation, including insulator activity, gene activation and repression, genomic imprinting and tumor suppressor [4].

Study design, data and methods

Our study primarily relies on recent data published by the ENCODE consortium [5]. The general idea is to use machine learning algorithms to build models that predict site occupancy at predefined target sites. We used two types of candidate target site list: (a) Predicted sites from a whole genome scan with a position weight matrix (PWM) and (b) 10 cell type-specific peak lists published by ENCODE. For both types of candidate sites we expressed cell-type specific CTCF occupancy as the number of ChIP-Seq tags within a window of 200 bp around the site. In parallel, we collected for each site in each lists a number of associated predicted and experimental features (Table 1). We then applied machine learning algorithms to predict site occupancy from different subsets of associated features. Note that experimental features such as histone modifications or DNase I hypersensitivity were evaluated in two different ways: (a) by the total number of tag counts in a window around the site and (b) by computing a "shape-score" reflecting the similarity of the tag distribution around a particular site with the average tag distribution as seen in an aggregate plot. We applied machine learning in a binary class-prediction framework and by regression analysis. For class prediction the candidate site lists were first split into high and low occupancy classes. Support vector machines (SVM) combined with recursive feature selection performed better than random forests (RF) was used for class prediction, support vector regression (SVR) was used for model training with quantitative site occupancy data. The performance was evaluated by 10-fold cross-validation, except in the cases where data from one cell type were used to predict the results from another cell type. Performance was expressed as a Pearson correlation coefficient between predicted and experimentally determined site occupancy (Fig. 1A).

Summary of results and selected examples

Rad21, TFBS-score and DGF are the most important features that contributed significantly to the classification followed by histone marks. It was known before that CTCF associates with Rad21 in the so-called CTCF/cohesion complex. The very good performance of Rad21 confirms the previous reports that CTCF acts in close coordination with this protein. Other sequence and structural features showed relatively low importance and didn't contribute significantly to the classification accuracy when considering TBFS-score alone, however, they performed better in predicted CTCF sites than ENCODE peak list. In order to get an objective comparison between the two datasets we used top two features (showing maximum contribution) to build model on one cell line and prediction on other cell lines. In addition, five histone marks were also used for classification in five cell lines (limited due to dataset availability), where H3K4me1 a distal mark showed highest importance in classification. The two cancerous cell lines in the dataset K562 and HepG2 showed similar patterns among themselves but a distinct pattern when compared to other normal cell lines indicating that they has varied cell specific CTCF sites.

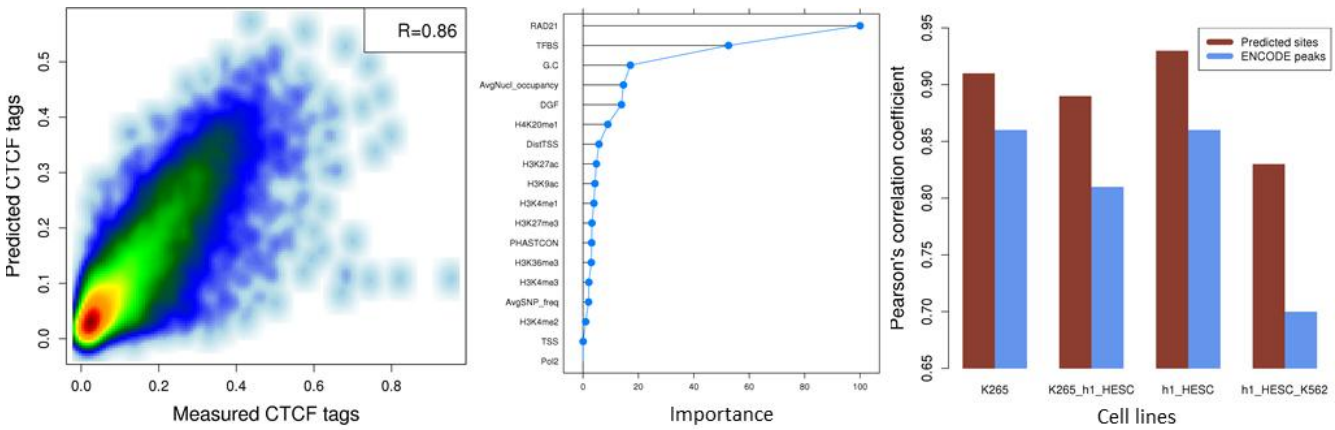

Figure1: A) SVR model was built on training dataset from K562 and used to predict CTCF tags on test dataset from K562 cell line. Graph shows the Pearson's correlation for measured and predicted CTCF sites on test dataset. B) Feature importance for regression on CTCF sites from ENCODE peak list for K562 cell line. C) SVR model was built on K562 and H1-hESC cell line for both dataset (ENCODE peak list and predicted CTCF sites), these models were then used to predict the CTCF tags on other cell line (Pearson's correlation coefficient shows the performance of model on cross-cell line prediction).

Table 1: Overview of features used to predict CTCF site occupancy

Feature type	**Feature description**	**Number**	**Data source/ Reference**
Sequence intrinsic	Mono-penta nucleotide frequency	1364	
	Nucleosome occupancy	1	[6]
	Structural parameters	10	[7]
	CTCF TFBS-score	1	JASPAR [8]
Evolution/Polymorphism	Avg PhastCons score	1	UCSC database
	SNP Frequency	1	dbSNP132
Tissue-specific experimental features (tag counts in window around site and shape based evaluation)	Distance to nearest TSS	1	ENSEMBL database
	DNase I	1	GEO series (GSE26328)
	Histone modifications	8	GEO series (GSE29611)
	PolII	1	GEO series (GSE32465)
	RAD21	1	GEO series (GSE32465)

To investigate the degree of tissue specificity we built models using TFBS-score, DGF, Rad21, average nucleosome occupancy/ base, average distance from TSS site and certain histone marks performed best in the prediction (Figure 1B). The results showed in general high correlation coefficient within and

across cell lines with certain notable exception (Figure 1C). Models built from fibroblast cell lines were good predictor of other fibroblast cell lines compared to other cell lines. The other cell line which showed consistently varied performance in cross-prediction experiment is K562, the cancer cell line considered in this study. We also note that tissue-specificity of trained models are better predictor of PWM predicted sites lists than experimental peak lists. This is not really surprising as the experimental peak list exclude those sites which totally unoccupied in given tissue.

References

[1] Valouev et al., 2008, Nat Methods, 5, 829-834
[2] Barski et al., 2007, Cell, 129, 823-836
[3] Neph at al., 2012, Nature, 489, 83-90
[4] Lee et al., 2012, J Biol Chem, 287, 30906-30913
[5] The ENCODE Project Consortium, 2012, Nature, 489, 57-74
[6] Kaplan et al., 2009, Nature, 458, 362-366
[7] Florquin et al., 2005, Nucleic Acid Res, 33, 4255-4264
[8] Portales-Casamar et al. 2010, Nucleic Acids Res, 38, D105–D110