Genomics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods Paper

Recognition of long-range enhancer-promoter interactions by adding genomic signatures of segmented regulatory regions

Zhen-Xing Feng, Qian-Zhong Li *

Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

ARTICLE INFO

Article history:
Received 7 January 2017
Received in revised form 16 May 2017
Accepted 31 May 2017
Available online xxxx

Keywords:
Enhancer-promoter interaction
Epigenetic
Nucleosome occupancy
Enhancer RNA
Random Forest

ABSTRACT

Enhancer-promoter interaction (EPI) is an important cis-regulatory mechanism in the regulation of tissue-specific gene expression. However, it still has limitation to precisely identity these interactions so far. In this paper, using diverse genomic features for various regulatory regions, we presented a computational approach to predict EPIs with improved accuracies. Meanwhile, we comprehensively studied more potential regulatory factors that are important to EPIs prediction, such as nucleosome occupancy, enhancer RNA; and found the cell line-specificity and region-specificity of the contributions of diverse regulatory signatures. By adding genomic signatures of segmented regulatory regions, our best accuracies of cross-validation test were about 11%–16% higher than the previous results, indicating the location-specificity of genomic signatures in a regulatory region for predicting EPIs. Additionally, more training samples and related features can provide reliable performances in new cell lines. Consequently, our study provided additional insights into the roles of diverse signature features for predicting long-range EPIs.

1. Introduction

Numerous DNA reactions in eukaryotic cell are regulated by the cooperativeness of diverse cis-regulatory elements, such as transcriptional promoters, enhancers and insulators [1,2]. Due to chromatin folding, three-dimensional chromatin organization brings distal enhancers in connection with target promoters in cis to regulate gene expression. In this way, the enhancers can directly interact with their target genes over tens of kilobases away genomic distances in three-dimensional space [3–5]. The long-range interactions can play a critical role in the regulation of tissue-specific gene expression [5,6] and may be linked to human diseases [7]. Generally, physical interactions between functional elements in genome are crucial for regulating gene transcription [8]; and an enhancer would contact with its target promoter(s) via a preformed architecture, and its modification [9,10] or the formation of other regulatory factors complexes [8,11]. In recent years, a number of high-throughput interaction capture techniques have made it possible to explore the spacial interactions, such as Fluorescence In Situ Hybridization (FISH) [12], Chromosome Conformation Capture (3C) [13], 4C (circular 3C) [14], 5C (3C-carboncopy) [15], Hi-C (3C variant) [16], paired-end tag sequencing (ChIA-PET) [17] and Targeted Chromatin Capture (T2C) [18]. These methods provided a detailed genome-wide information on the three-dimensional organization of the mammalian genome for cell ensembles. Recent works based on the above methods are helpful for understanding the long-range interactions mechanism of 3D chromatin organization, such as the roles of looping factors, gene regulatory elements and non-coding RNAs in defining specificity for EPIs [19], dynamics, accessibility, balance stability and flexibility in ensuring genome integrity and variation enabling gene expression/regulation [20], and disruptions of topological chromatin domains in causing pathogenic rewiring of gene-enhancer interactions [21]. Importantly, only sufficiently stable physical contact can result in a high contact probability between two loci, otherwise the pronounced fluctuations would effectively segregate them [22].

Based on above experimental data (e.g. 5C, Hi-C, T2C), it remains a challenging work to study the interactions between the enhancers and their far target promoter(s) in three-dimensional nuclear space. Especially with the impressive advancements in the field of high-throughput techniques, diverse genomic signatures were generated from different experimental platforms. For example, the ENCODE project [23] has generated many deep sequencing data profiles (e.g. >120 transcription

Abbreviations: EPI, enhancer-promoter interaction; E-P, enhancer-promoter; 5C, Carbon Copy Chromosome Capture Conformation; ChIA-PET, Chromatin Interaction Analysis by Paired-End Tag Sequencing; FISH, Fluorescence In Situ Hybridization; ChIP-seq, chromatin immunoprecipitation sequencing; T2C, Targeted Chromatin Capture; FAIRE-seq, Formaldehyde-Assisted Isolation of Regulatory Elements; eRNA, enhancer RNA; TFs, transcription factors; HMs, Histone modifications; TAD, topological associating domains; RPKM, the reads per kilobase of exon model per million mapped reads; GRO-seq, global nuclear run-on sequencing; MNase-seq, sequencing of micrococcal nuclease sensitive sites; RRBS, Reduced Representation Bisulfite Sequencing; CAGE, cap analysis of gene expression; HMM, Hidden Markov Model; GEO, Gene Expression Omnibus; TSSs, transcription start sites; AUROC, area under the receiver operating characteristic curve; AUPR, area under the precision-recall curve.

* Corresponding author.
E-mail address: qzli@imu.edu.cn (Q.-Z. Li).

http://dx.doi.org/10.1016/j.ygeno.2017.05.009
0888-7543/

factors (TFs), >11 Histone modifications (HMs), the cap analysis of gene expression (CAGE), Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq)). These data have made it practicable to study the relationships of diverse genomic signatures on different genomic scales [24]. In past years, by using these related genomic marks, many efforts have been made to develop the computational methods to predict EPIs. For example, a logistic regression classifier was proposed to predict EPIs by using correlations between profiles of chromatin state, gene expression, regulatory motif enrichment and regulator expression [25]. Then, based on the H3K4me1 signal, CTCF binding sites and RNA-seq data, the method of predicting specific tissue interactions of genes and enhancers (PreSTIGE) was proposed to delineate EPIs [26]. Afterward, with four genomic features (i.e. correlation between enhancer and target promoter activity profile, correlation between TFs and target promoter, the coevolution of enhancer and target promoter, and distance constraint between enhancer and target promoter), the integrated method for predicting enhancer targets (IM-PET) was proposed to train a classifier for predicting EPIs [27]. In 2015, a high-throughput identification pipeline called HIPPIE was implemented for predicting EPIs [28]. However, due to the cell line-specificity of EPIs and the difficult experimental verification issues in a new cell type or time point, a systematic framework of regulatory Interaction prediction for promoters and long-range enhancers (RIPPLE) using genomic features (including 8HMs, 15TFs, Dnase I and RNA-seq) was developed to predict EPIs with a high prediction performance [29].

As demonstrated by a series of recent publications [30–33] in compliance with Chou's 5-step rule [34], to establish a really useful statistical predictor for a biological system, we should follow the following five guidelines: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly webserver for the predictor that is accessible to the public. In this study, based on the previous work [29], we aim to develop a more reliable and efficient predictive model to identify the true EPIs from the possible interactions pairs, and further detect more potentially influential signatures for the prediction and their location, distribution and correlation properties. To this end, we extracted features for enhancer, promoter and loop regions from diverse genomic signatures; the signatures contained TFs, HMs, Dnase I, DNA methylation, RNA-seq, nucleosome occupancy, etc. Then, we integrated these features and established a reliable classifier to predict the EPIs with higher accuracies. Meanwhile, we not only comprehensively studied previously reported elements but also found the important roles of several other factors (e.g. enhancer RNA, nucleosome occupancy) for predicting EPIs. From our analysis, we found the distinct distribution differences of the influential features between interacting and non-interacting sets. Finally, our further improved results suggested the great contributions of genomic features in the segmented regulatory regions to the EPIs prediction.

2. Materials and methods

2.1. Dataset of EPI pairs

In this study, the dataset of EPI pairs for Gm12878, H1hesc, Hela and K562 cell lines were taken from the Roy et al.'s work [29]. The positive dataset was interrogated in 1% of the human genome (ENCODE pilot project regions) from the 5C experiments [15], and contained 877, 765, 476 and 337 EPI pairs in K562, Hela, Gm12878 and H1hesc cell lines, respectively. Because an enhancer has a great tendency to interact with target promoters depending on their distances [27,28], Roy et al. derived the justified negative datasets by controlling the distances of enhancers and promoters in the same distribution with the positive samples (Fig. 1c) [29]. Moreover, the number of E-P pairs in the negative set was the same as the positive one in the corresponding cell line. For further analysis, we counted the number of shared E-P pairs across multiple cell lines (Fig. 1a).

2.2. The extraction of signature features

2.2.1. The TFs and FAIRE-seq

Binding to cis-regulatory elements of downstream target genes, different types of TFs play critical roles in tissue specific gene expression. In our work, we used the available chromatin immunoprecipitation sequencing (ChIP-seq) assay data of TFs in each cell line from the ENCODE project. The files of ChIP-seq peak were download from ENCODE ChIP-seq Experiment Matrix hg19 (http://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html). FAIRE-seq peak files were also from the website. The names of different type of TFs in four cell lines were listed in Supplementary Table 1.

Motivated by the previous works [35,36], we took the average intensity of TFs to generate features for a regulatory region (i.e. enhancer, promoter and loop region). In our work, the average intensity (A_{ij}) was defined as a measure of the binding strength of TF j for a regulatory region of i-th E-P pair. We computed the average intensity of the ChIP-seq peaks of TF j depending upon whether a particular peak had ≥1 bp overlap with a regulatory region of i-th E-P pair.

$$A_{ij} = \sum_{k=1}^{K} g_j(k)/L_i \tag{1}$$

where $k(k=1,2,3,..,K)$ iterates over all peaks of TF j, and the peaks have ≥1 bp overlap with a regulatory region of i-th E-P pair, $g_j(k)$ is the height of peak k of TF j, L_i is the length of a region in i-th E-P pair. Thus, for each type of TF, we obtained a 1-dimensional feature vector for enhancer, promoter and loop region, respectively. The binding intensity (A_{ij}) of FAIRE-seq had the same computational method with TFs.

2.2.2. HMs and Dnase I

Correlated well to establish EPIs, the HMs have been successfully applied to the EPIs prediction [26–29]. In this paper, we used ChIP-seq assay data of 11 HMs (H2az, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3 and H4K20me1) shared in four cell lines from the ENCODE project, and downloaded the bam files of HMs and Dnase I from ENCODE ChIP-seq Experiment Matrix hg19 (http://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html). Then we converted the files of bam format to corresponding files of bed format using the BEDTools [37].

Considering that the ChIP-seq patterns for HMs do not exhibit the well-defined 'peaks' present in TF-binding data, we calculated the tag density of 11 HMs for the enhancer, promoter and loop regions, and quantified the tag density in the computational method of normalized RPKM definition [38]. The tag density (D_{ij}) of the HM j for a regulatory region of i-th E-P pair was defined as follows.

$$D_{ij} = 10^9 \times n_{ij}/(L_i \times N_j) \tag{2}$$

where n_{ij} is the total tag count of HM j that was located in a region of i-th E-P pair, L_i is the length of the region in i-th E-P pair, N_j is the total tag count of HM j in the measurement. Thus, for each type of HM, we obtained a 1-dimensional feature vector for the enhancer, promoter and loop region, respectively. In addition, we averaged the tag density (D_{ij}) for the two biological replicates of HM j in enhancer, promoter and loop regions, respectively. The computational method of tag density D_{ij} of Dnase I was the same with HMs.

Please cite this article as: Z.-X. Feng, Q.-Z. Li, Genomics (2017), http://dx.doi.org/10.1016/j.ygeno.2017.05.009