Promoter-Enhancer association study

BY YIZHE ZHANG, YUNKE MAI

Email: yz196@duke.edu, yunke.mai@duke.edu

1 Motivation

Understanding enhancer-promoter interaction can reveal how regulatory networks work. ChIA-PET experiment was devised recently to identify unique, functional chromatin interactions between distal and proximal regulatory elements (REs) and the promoters of the genes they interact with, which allows for a genome-wide modeling of promoter-enhancer(PE) interaction .

LDA may enable such inference by let us find latent "regulatory element topics". Each "regulatory element topics" can be a bunch of regulatroy elements that work together. As an analogy, the promoter or enhancer region can be seen as documents. The regulatory elements on them can be seen occurences of words.

Promoter-Enhancer interaction	Document citation
Promoter/Enhancer	Documents
TFBS/histone markers	Words
Promoter Enhancer interaction	Citation between documents

Table 1. Analogy of document-citation scenario

In order to inference the interactions we use Relational Topic Modeling (RTM). It provide a framework to integrate link/reference information with document information. However, in standard RTM documents share same vocabulary, while the topics found in enhancer and promoter may use different set of vocabulary. Thus, we hope to integrate RTM and bilingual topic modeling to fit our data.

This project will first examine how PE interaction can be explained by REs on both promoter and enhancer, using a generalized linear model. The model significance will be assessed and model comparison will be conducted. Then, based on the GLM result, a topic modeling framework will be used.

2 Method

2.1 Generalized linear model

We will first focusing on fitting a generalized linear models (binomial family) including all interaction terms of REs (counts or identity) between promoter and enhancer regions:

For data point i:

$$\begin{aligned} \operatorname{logit} \pi_i &= \alpha + \sum_k \sum_t \beta_{kt} P_{ik} E_{it} + \beta_d \log(d) \\ p(y_i = 1 | P_i, E_i) &= p_i \end{aligned}$$

Where $P_{\cdot k}, E_{\cdot k}$ are the counts of kth RE on genomic region. d is the distance between two region.

Another possible model is to first exact topics of each region (using LDA), then apply generalized linear model on the topic level, based on the topic assignments.

2.2 Relational Topic Modeling

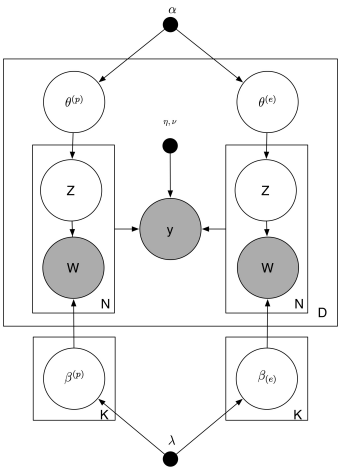

Figure 1. Model design

The generative story will be

1. Generate promoter's topic distribution $\theta_i^{(p)}$ for each promoter i: $\theta_i^{(p)} \sim \mathrm{Dir}(\alpha)$
2. Generate topic assignment for each word: $Z_{i,j}^{(p)} \sim \mathrm{Multi}\left(\theta_i^{(p)}\right)$
3. Generate each topic's word distribution: $\beta_k^{(p)} \sim \mathrm{Dir}(\lambda)$
4. Generate each word: $W_{i,j}^{(p)} \sim \mathrm{Multi}\left(\beta_{Z_{i,j}}^{(p)}\right)$
5. Similarly, generate the corpus for enhancer.
6. Generate link with exponential link: $p(y_{p,e}=1) = \exp\left\{\eta^T \bar{Z}^{(p)} \circ \bar{Z}^{(e)} + \nu\right\}$

Where $\circ$ denote elementwise product. η, ν are contrained to ensure probability$\in[0, 1]$. Intuitively, documents with similar topic distribution will have higher probability of link.

The reason we use exponential link is for computational conveniency, as well as not losing accuracy[3]. However we can also try links like logit and probit. Gibbs sample is possible for exponential link.

2.3 Data preparation

We first preprocess the rawdata. First the promoters/enhancers were labeled by the existence/absence of TSS (transcription start site). The result was filtered so that only promoter-enhancer pair was collected. All regions have no intersection with each other.

For each ChIP-seq peak region of certain RE, if it overlaps with any promoter/enhancer, we add this RE words into the promoter/enhancer once.

The distance feature between promoter and enhancer is treated as a compound covariate: an indicator of whether they are on the same chromesome and a real value of the distance between the midpoint of each region's genomic coordinates (if they fall into the same chromesome).

2.4 Generate negative dataset

We assume the interactions between REs (or RE topics) can partly explain the spatial interactions we observed between Promoter and Enhancer. However the only data we have is the known links (via RNA polymerase II) between genomic segments. In another words, we only have positive dataset. In order to generate a negative datasets represent the null, we use permutation. In specific, we arbitarily link a promoter with an enhancer, and make sure they are not originally paired.

The hypothesis we want to test is like: "Promoter-enhancer link is influenced by REs interaction". Thus permuting over the order disrupt the REs interaction, and can represent the null to some extents. However, this approach may admittably suffer from several flaws. First, the links between reordered promoters and enhancers may not be exactly zero. Unseen doesn't means non-link. Instead there may be an average(background) level of probability. Second, other covariates like distance need to be conditioned on, however there may be no easy way to do this. What we can do is to select the alternative promoter/enhancer nearby the original promoter/enhancer.

The size of negative dataset may also influence the result. According to our observation, the impact is quite large: the prediction of link probability on test set tend to biase toward the proportion of "link" response in the training set. We set the negative set equal to positive set to make the prediction roughly centered at half.

3 Result

3.1 Generalized liner model

We treat the links as response. The data was divided into ten partitions and nine was used for training, one for testing.

For the MCF7 dataset. We compare a RE interaction model (the predictors are all interactions between REs) with a topic interaction model (the predictors are all interactions between infered topics). We use 15 topics.

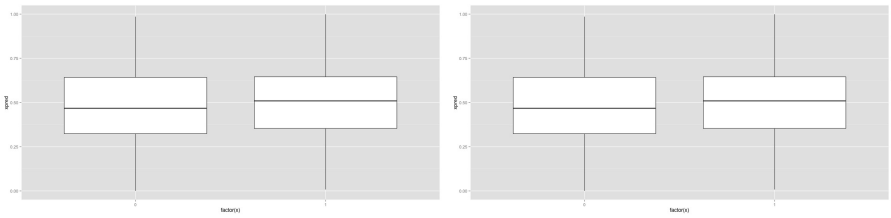

Figure 2. Predict link probability (left panel: REs as indicator right panel: REs as count)

p-value	Promoter	Enhancer
0.0006	tcf12	egr1
0.0002	foxm1	e2f1
3.4e-05	tead4	gabpa
4.8e-05	tcf4	tcf4
0.0009	tcf4	nr2f2
0.0008	tcf4	rad21

Table 2. Significant RE interactions (p-value<0.001)

The result indicate REs information alone can predict the links. For all 26^2 interactions 6 has pvalue<0.001. We see Tcf4 may be a pivot factor of predicting PE links.

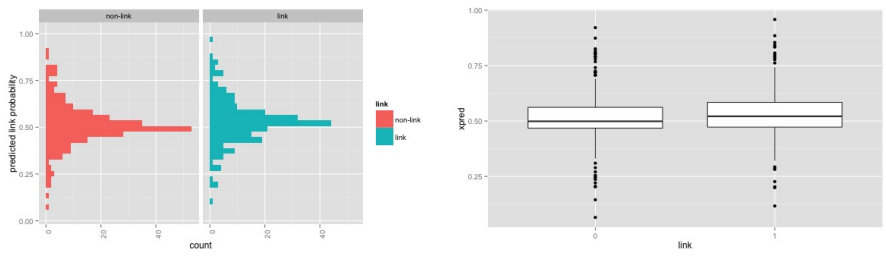

Figure 3. Topic level link prediction

By packing the information into topics, and treating them as cooperative RE groups, we still maintain slight predictive power. (The difference is slight but robust for ten repeat trials)

The ANOVA test comparing REs models and topic model show no significant improvement.

We further consider add distance between the pair into prediction, which result in two additional covariates: whether in same chromesome and log distance between two region (infinity for regions not in same chromesome)

We compare the result between using merely distance and using distance with RE topics. After adding RE topics, a subset of covariates quasi-complete separate the responses. The estimation of coefficients become large. Thus, we put norm-1 regularization (LASSO) on both of the coefficients (using "glmnet" package).

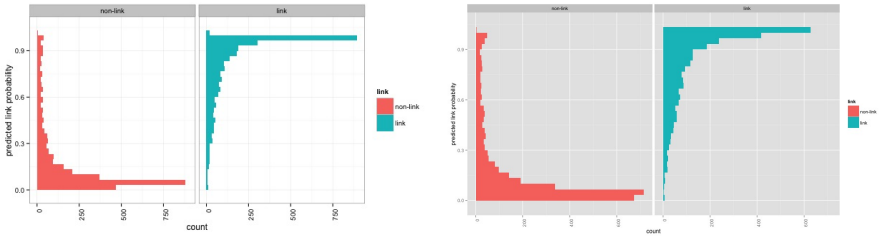

Figure 4. left: distance alone right: distance+RE topics

For K563 dataset, the full interaction model will be too large to fit (111^2 predictors fitting 50,000 PE pairs). So we only tried using topics to fit glm.

3.2 Bilingual topic modeling

First we test whether promoter and enhancer share same RE topics. We use perplexity to assess how well the infered topics can intepret new documents. Small perplexity indicate better prediction.

$$\text{perplexity} = \exp\left(-\frac{\log(W_d)}{N_d}\right)$$

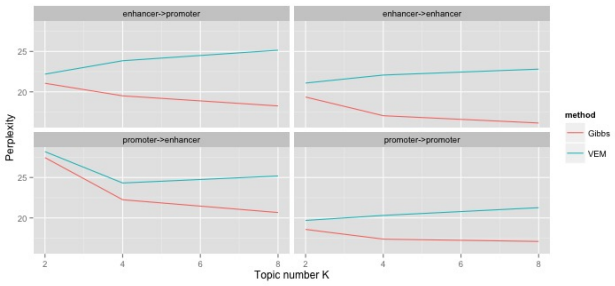

Figure 5. enhancer and promoter use different topics.

If we use topics learned from enhancer/promoter corpus to predict promoter/enhancer, the perplexity is not good as using enhancer/promoter corpus to predict enhancer/promoter, which indicate separate modeling for enhancer and promoter are necessary.

4 Data

- Mouse ChIA-PET data from Yubo Zhang (Nature 2014)[1]
- Human ChIA-PET data from Guojiang Li (Cell 2012)[2]

- 26 human TFs ChIP-seq from ENCODE project (MCF7)
- 111 human TFs, histone markers and Dnase signals. (K562)
- TSS: SwitchDB TSS track.

5 Reference

1. Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature 504, 306–310 (2014).

2. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98 (2012).

3. Chang, J. & Blei, D. M. Hierarchical Relational Models for Document Networks. The Annals of Applied Statistics 4, 124–150 (2010).