Predicting enhancer-promoter interaction using genomic sequence features

Shashank Singh, Yang Yang, Ruochi Zhang, Barnabás Póczos, Jian Ma

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Introduction

In the human genome, a large number of distal enhancers regulate target genes through proximal promoters by forming enhancer-promoter interactions (EPIs). Recent high-throughput chromatin interaction mapping methods allow us to recognize potential EPIs, but it is largely unknown if the sequence level features are sufficient to build a predictive model for EPIs.

Research Questions

- Are there sequence-level EPI determinants?
- If so...
  - What are they?
  - Are they sufficient to be used to predict EPIs?
  - How consistent are they across cell lines?

Data

1. Active enhancers and promoters identified from ENCODE [1] and Roadmap Epigenomics [3] annotations, in each of 6 cell lines.
2. As in [2], EP pairs were annotated as positive (interacting) or negative (non-interacting) using cell-line-specific genome-wide chromatin contact measurements based on Hi-C [4].
3. 20 negative pairs sampled per positive pair
   - positive/negative pairs were constrained to have similar distributions of enhancer-promoter distance
4. Thus, data are heavily imbalanced ($> 95\%$ negative), in accordance with the fact that most enhancer/promoter pairs do not interact.

References

[1] ENCODE Project Consortium. *Nature*, 489(7414):57-74, 2012.

[2] Whalen et al. *Nature Genetics*, 48(5):488-496, 2016.

[3] Roadmap Epigenomics Consortium. *Nature*, 518(7539):317-330, 2014.

[4] Rao et al. *Cell*, 159(7):1665-1680, 2014.

[5] Chen and Guestrin. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.

Acknowledgements

NSF Graduate Research Fellowship DGE-1252522 (to S.S.); NIH grants HG007352, DK107965, and CA182360, and NSF grants 1054309 and 1262575 (to J.M.).

Motif/Embedding Model (PEP)

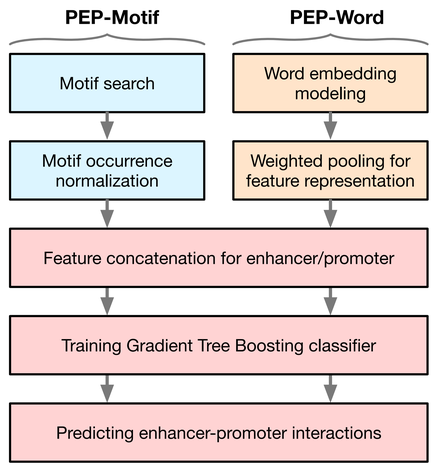


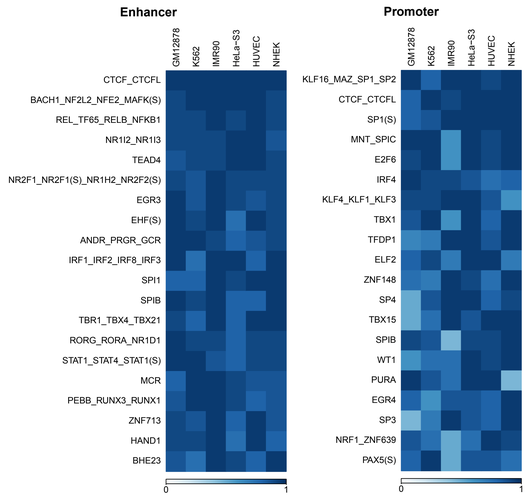
Figure 1: Diagram of PEP pipeline.

PEP learns a gradient tree boosting model [5] using two sets of features:

- PEP-Motif finds known transcription factor binding site (TFBS) patterns.
- PEP-Word learns a word embedding model to obtain continuous distributed feature representation of sequences.

Sequence Features from PEP

We ranked clustered TF features from PEP-Motif based on feature importance estimated using XGBoost.



Deep Learning Model (SPEID)

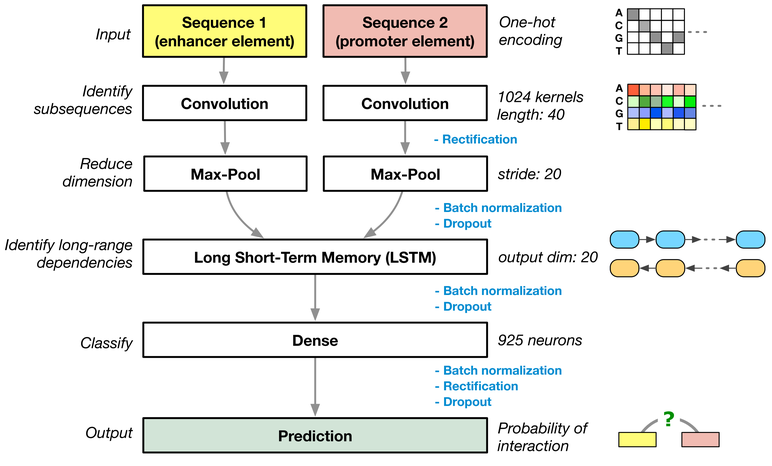


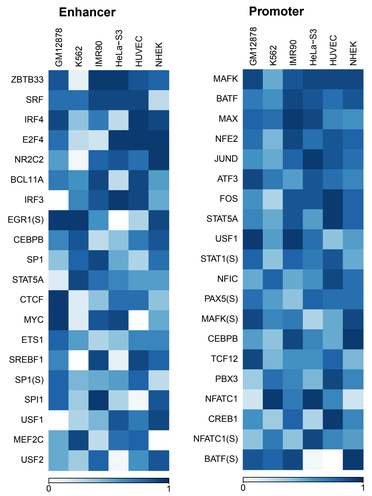
Figure 2: Diagram of SPEID network.

SPEID has three main sequential layers:

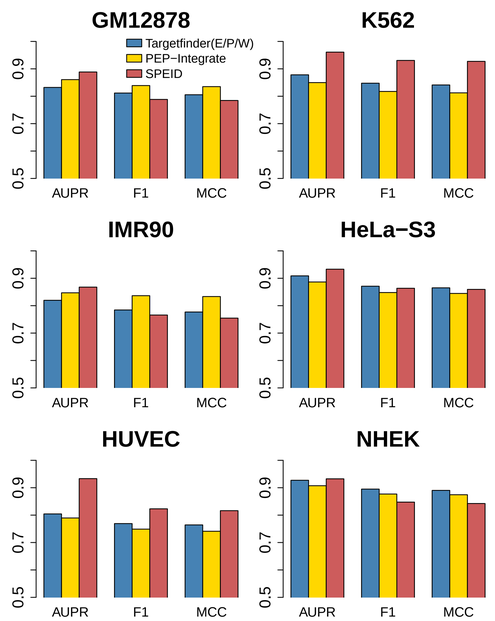
- Parallel convolution/pooling layers learn to extract short (40bp) sequence features from both inputs.
- LSTM layer learns to identify interactions between sequence features and between inputs.
- Dense layer predicts EPI from these high-level features.

Sequence Features from SPEID

We ranked importance of TF's in SPEID using *in silico mutagenesis* (replacing TFBS with noise and measuring impact on prediction performance).



Prediction Results



Conclusions

Main Result

Proposed methods achieve state-of-the-art EPIs prediction performance using only DNA sequence-based features. Thus, sequences encode the vital mechanisms mediating EPIs.

- Recent machine learning models and representations for complex features, such as deep networks and word embeddings, can help extract crucial predictive information directly from genetic sequences.
- Sequence-based prediction models can identify sequence features predictive of EPI.

Carnegie Mellon University