TF-Finder: Predicting the Underlying Transcription Factors in Genomic Sequencing Data

Diwakar Ganesan and David Wu

Summary

Motivation: New techniques in biotechnology allow researchers to determine the set of genomic regions bound by transcription factors (TFs). However, there is no way to identify the specific transcription factors involved in such sets. Being able to do this would allow the identification of potential drug targets in a variety of diseases.

Inputs: PRISM predicted TF binding sites and a set of genomic regions of interest to decompose into TFs.

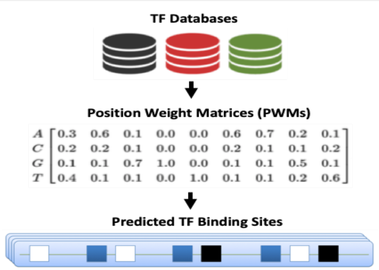
Outputs: Ranked list of TFs that are most likely responsible for interacting with the set of genomic regions of interest.

Background

Transcription factors (TFs) are a ~2,000 member subclass of proteins that are critical in regulating protein production. They are spatial-temporal switches, binding to specific DNA sequences to either turn on/off transcriptional machinery. Misregulation of transcription can lead to devastating diseases such as cancer and heart disease.

An ontology is systematic way to document both normal and diseased phenotypes, or observable characteristics of a person. Within each ontology are a set of **ontology terms**, such as "coughing" and "abnormal blood clotting."

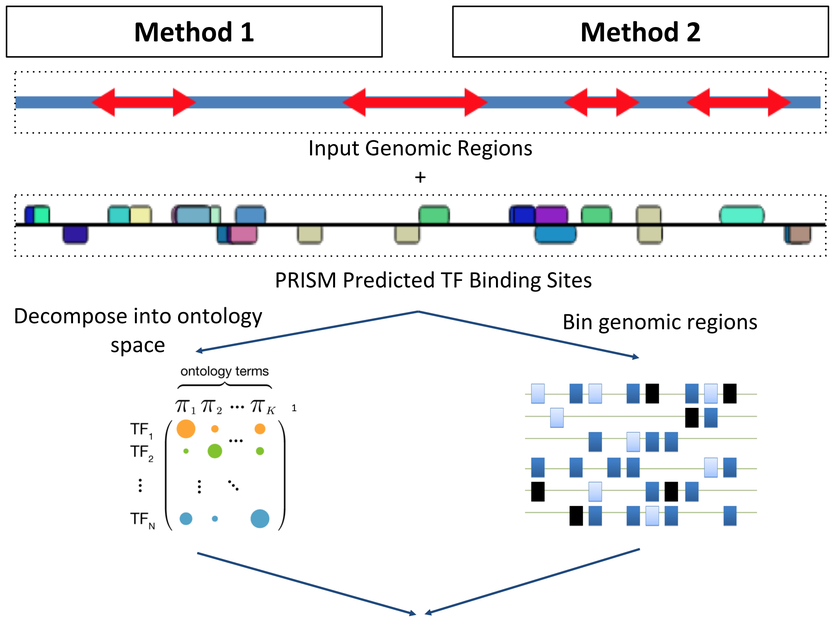
Data



Transcription Factor Binding Site Prediction (left) We fed TF position weight matrices (PWMs) to PRISM for high quality prediction of TF binding sites. PWMs were scraped from public TF databases and manually curated for quality.

Evaluation Set 49 sets of genomic regions (passing quality control) experimentally determined to be associated with TFs in our library were downloaded from the ChIP-Atlas database for use in an evaluation set. The GREAT computational pipeline was used to determine genomic region set quality.

Methods



Results

We produced rankings using a variety of ontologies for Method 1, and used 1kb genomic bin size for Method 2. Best results shown below.

Method 1: GOBiological

	Baseline	PCA	NMF
Top 10	0.204	0.04	0.224
Top 25	0.244	0.04	0.3469
Top 50	0.367	0.04	0.346

Method 2

	Baseline	PCA	NMF
Top 10	0.632	0.18	.55
Top 25	0.710	0.51	.69
Top 50	0.816	0.69	.73

Discussion

Our data give rise to two general results. First, ontologies are not a good way of representing genomic regions -- we observed genomic binning gave much better ranking performance. Second, we found NMF is a better way of compressing genomic data than PCA. NMF outperformed PCA in both sets of experiments, which indicates exploiting the nonnegative structure of the data is important. Surprisingly, dimensionality reduction did not yield better results in the 1kb genomic bin experiments, meaning we likely need more information on transcription factor biology and function, to better generate a biologically-consistent lower dimensional representation of the data.

Future Work

1. Investigate better genomic region encoding schemes that represent biological phenomena
2. Investigate PCA/NMF with sparsity constraints enforced
   a. This follows from the assumption that only small subsets of the genome are involved with most protein expression
3. Test our pipeline on novel genomic region sets associated with disease, and empirically verify predicted TF-genomic region relationships

References

[1] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," Cell, vol. 172, no. 4, pp. 650–665, 2018.
[2] A. C. Joerger and A. R. Fersht, "The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches," Annual review of biochemistry, vol. 85, pp. 375–404, 2016. [3] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, "Great improves functional interpretation of cis-regulatory regions," Nature biotechnology, vol. 28, no. 5, p. 495, 2010.
[4] Y. Tanigawa, J. Li, J. M. Justesen, H. Horn, M. Aguirre, C. DeBoever, C. Chang, B. Narasimhan, K. Lage, T. Hastie, et al., "Components of genetic associations across 2,138 phenotypes in the uk biobank highlight adipocyte biology," Nature communications, vol. 10, no. 1, pp. 1–14, 2019.
[5] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, "Jaspar: an openaccess database for eukaryotic transcription factor binding profiles," Nucleic acids research, vol. 32, no. suppl_1, pp. D91–D94, 2004.
[6] D. E. Newburger and M. L. Bulyk, "Uniprobe: an online database of protein binding microarray data on protein–dna interactions," Nucleic acids research, vol. 37, no. suppl_1, pp. D77–D82, 2008.
[7] I. V. Kulakovskiy, Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, I. E. Vorontsov, V. B. Bajic, and V. J. Makeev, "Hocomoco: a comprehensive collection of human transcription factor binding sites models," Nucleic acids research, vol. 41, no. D1, pp. D195–D202, 2012.
[8] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, et al., "TransfacR : transcriptional regulation, from patterns to profiles," Nucleic acids research, vol. 31, no. 1, pp. 374–378, 2003.
[9] A. M. Wenger, S. L. Clarke, H. Guturu, J. Chen, B. T. Schaar, C. Y. McLean, and G. Bejerano, "Prism offers a comprehensive genomic approach to transcription factor function prediction," Genome research, vol. 23, no. 5, pp. 889–904, 2013.
[10] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al., "Ncbi geo: archive for functional genomics data sets—update," Nucleic acids research, vol. 41, no. D1, pp. D991–D995, 2012.

1. https://www.biorxiv.org/content/biorxiv/early/2019/08/20/730200.full.pdf