CS230

Predicting cell type specific functional consequences of non-coding variation using deep learning for CS230-Winter 2019

Santosh Kumar[*]
Department of Computer Science
Stanford University
ksantosh@stanford.edu

Sangjukta Kashyap [†]
Department of Computer Science
Stanford University
sanju321@stanford.edu

Abstract

Predicting the functional consequences of genetic variants in non-coding regions is a challenging problem. We used here a deep learning approach, to jointly utilize experimentally confirmed regulatory variants (labeled variants), unlabeled variants genome-wide, and more than a thousand cell/tissue type specific epigenetic annotations to predict functional consequences of non-coding variants. Through the application to several experimental datasets, we demonstrate that the proposed method gets very good prediction accuracy,

1 Introduction

Determining the functional consequences of genetic variants is a difficult problem in human genetics. Our understanding of the genetic code and splicing enables us to identify variants that are likely functional in protein-coding regions, but accurately predicting the functional effects of variants in non-coding regions is much more difficult1. Multiple lines of evidence support an important functional role for variants in non-coding regions. For example, comparative genomic studies show that most of the mammalian conserved and recently adapted regions reside in the non-coding part of the genome. In addition, genome-wide association studies (GWAS) have identified a large number of non-coding variants that are likely to be involved in both genetic and epigenetic gene regulation in a highly context-specific manner2. Therefore, accurately predicting both organism level and cell type/tissue-specific functional consequences of non-coding variation is of great interest.

2 Related work

There are several possible approaches to predict the functional effects of genetic variants3. In the experimental approach (e.g. massively parallel reporter assays (MPRAs), CRISPR/Cas9- mediated in situ saturating mutagenesis), the functional effect of a variant is measured by evaluating the phenotypic consequence of the corresponding sequence alteration (e.g. by measuring the impact of individual alleles on gene expression in a particular context)4–6. This is considered the gold-standard approach, but it is quite laborious to perform in a comprehensive manner for large sets of genetic variants. More often, functional effects are derived using alternative approaches. One commonly used method is based on an evolutionary perspective, whereby functional effects are assessed by the

[*]https://www.linkedin.com/in/kumsantosh/
[†]https://www.linkedin.com/in/sangjukta-kashyap-99901b5/

CS230: Deep Learning, Winter 2019, Stanford University, CA.

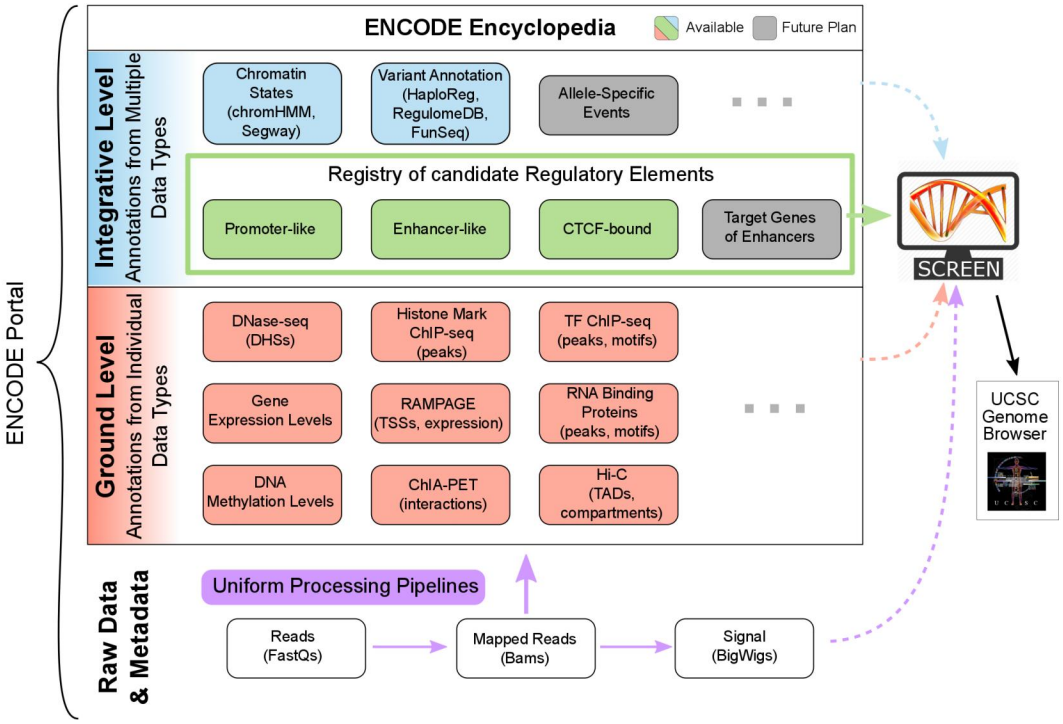
extent of evolutionary conservation at the position of interest. The classical evolutionary approach relies on accurate multispecies alignment, which makes it challenging to identify certain functional elements, such as elements constrained only within the human species, although several methods have been recently proposed to identify primate- or human-specific conserved elements7–9. Evolutionary approaches also pose an additional challenge, namely they cannot reveal the relevant cell type or tissue. Another popular approach is the biochemical approach, based on ChIP and/or DNase I hypersensitivity assays, with the caveat that such biochemical signatures can occur stochastically, and hence do not completely imply functionality. Therefore, depending on the approach, functional effect can have different meanings in different contexts. This creates challenges for meaningful comparisons among the different approaches.

3 Dataset and Features

The dataset we got from Dr. He from Stanford University is ENCODE. The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODE results from 2007 and later are available from the ENCODE Project Portal, encodeproject.org. This covers data generated during the two production phases 2007-2012 and 2013-present. The ENCODE Project Portal also hosts additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.

UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. All ENCODE data at UCSC are freely available for download and analysis.



Label: LabelData$_{XX}$.txt, "$Label$"$column$

Features: LabelData$_{XX}$.txt, $8 * 127 columns after$"$Label$"

Features example:

DNase-E001

where,

DNase is feature and

E001 is Cell type/Tissue type

Below is the table

	chr	pos	rs	Label	DNase-E001	H3K27ac-E0	H3K27me3-I	H3K36me3-I	H3K4me1-E(	H3K4me3-E(	...	H3K9ac-E12	H3K9me3-E:	DNase-E129	H3K27ac-E1	H3K27me3-I	H3K36me3-I	H3K4me1-E:	H3K4me3-E:	H3K9ac-E12	H3K9me3-E129
0	chr10	101980135	chr10:10198(	0	0.47	0.4	0.28	4.36	0.42	0.25	...	0.43	0.34	0.5	0.19	0.13	1.84	0.33	0.11	0.26	0.15
1	chr10	102010516	chr10:10201(	0	0.45	0.43	0.34	3.75	0.42	0.3	...	0.32	0.37	0.45	0.36	0.33	4.08	0.4	0.31	0.4	0.38
2	chr10	102012645	chr10:10201:	0	0.36	0.29	0.28	1.12	0.29	0.28	...	0.33	0.28	0.48	0.31	0.23	1.38	0.26	0.3	0.36	0.3
3	chr10	102031373	chr10:10203:	0	0.44	0.33	0.35	0.69	0.56	0.35	...	0.36	0.58	0.48	0.2	0.15	1.31	0.28	0.2	0.24	1.08
4	chr10	102265445	chr10:10226(	0	0.39	0.07	0.69	0.38	0.12	0.05	...	0.13	0.1	0.58	0.59	0.45	1.84	0.38	0.37	0.49	0.33

4 Methods

We used deep learning model using labels and features listed above to train to train different models.

Compare the method with existing alternatives. We can start with E116. - try all 1000 features or 8 E116-specific features

Methods Used:

LogisticRegression
DecisionTreeClassifier
KNeighborsClassifier
BernoulliNB
LinearDiscriminantAnalysis
GaussianNB
RidgeClassifier
SGDClassifier
Support Vector Machines Non-regularized NN model
NN model with L2 Regularization
NN Model with Dropout

5 Experiments/Results/Discussion

Github link for code: https://github.com/ksantosh321/CS230W2019

1) Comparing the results of each classifier

Test Results		
Classifiers	Mean ROC (AUC Value)	Accuracy
LogisticRegression	0.72	95%
DecisionTreeClassifier	0.57	94%
KNeighborsClassifier	0.57	95%
BernoulliNB	0.5	94%
LinearDiscriminantAnalysis	0.69	96%
GaussianNB	0.68	95%
RidgeClassifier	0.55	97%
SGDClassifier	0.72	95%
Support Vector Machines	0.65	95%

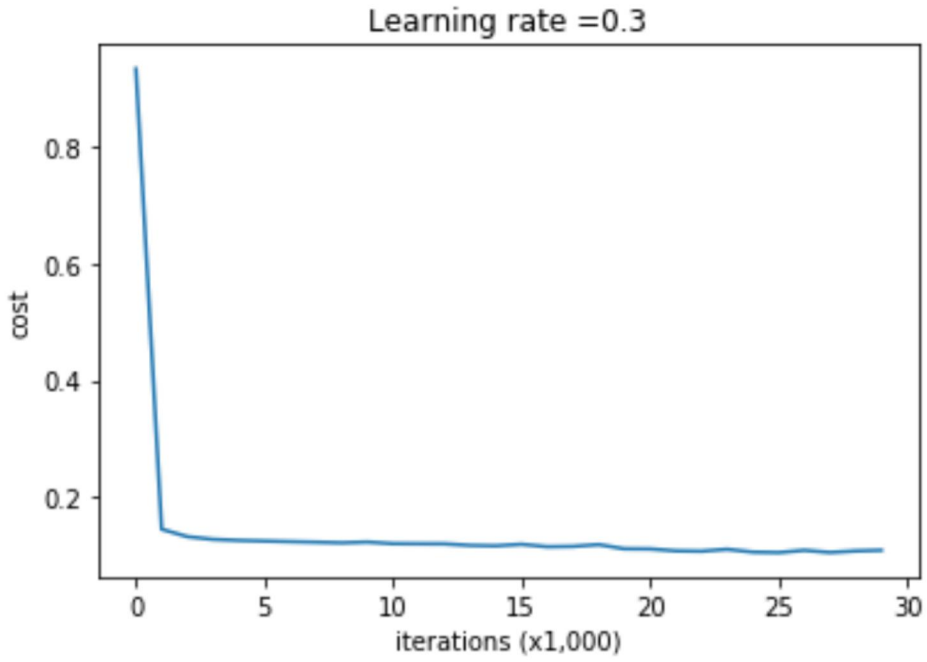
https://github.com/ksantosh321/CS230W2019/blob/master/ROC Curve Classifiers V2 with Train and Test .ipynb

2) Other NN Models

Test Results	
Classifiers	Accuracy
Non-regularized NN model	96.8%
NN model with L2 Regularization	96.3%
NN Model with Dropout	97.1%

NN Model with Dropout

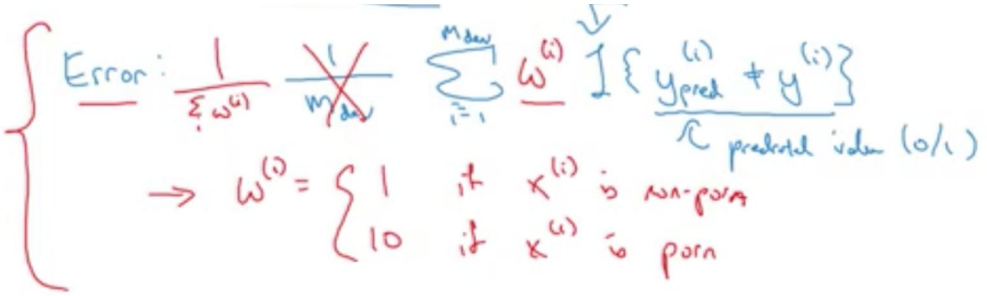
```
Cost after iteration 1000: 0.12018047107819292
Cost after iteration 2000: 0.11106862917607915
```




```
On the train set:
Accuracy: 0.970911651589838
On the test set:
Accuracy: 0.9711871750433275
```

6 Conclusion/Future Work

The current dataset has only 3% as label 1, and we tried weighting as advised to calculate the cost.



Most of classfier and model we trained and tested are getting high accuracy and we have scope of making ROC(AUC) curve better.

We also tried convolution but we didn't much improvement ROC(AUC) curve.
Next step would be to get bigger data set and try Convolution and other Neural network.

7 Contributions

We gratefully acknowledge support:

1) Our Mentor Hoormazd Rezaei

https://www.linkedin.com/in/hoormazd-rezaei/

2) Dr. He for sharing the dataset with us and guiding us during this project

https://profiles.stanford.edu/zihuai-he

References

1. Ritchie, G. R., Dunham, I., Zeggini, E. Flicek, P. Functional annotation of noncoding sequence variants. Nat. Methods 11, 294–296 (2014).

2. Altshuler, D., Daly, M. J. Lander, E. S. Genetic mapping in human disease. Science 322, 881–888 (2008).

3. Kellis, M. et al. Defining functional DNA elements in the human genome. Proc. Natl Acad. Sci. USA 111, 6131–6138 (2014).

4. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell 165, 1519–1529 (2016).

5. Kheradpour, P. et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 23, 800–811 (2013).

6. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819–823 (2013).

7. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. Nat. Genet. 46, 944–950 (2014).

8. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 9, e1003709 (2013).

9. Petrovski, S. et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. PLoS Genet. 11, e1005492 (2015).

10. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

11. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. 28, 1045–1048 (2010).

12. Martens, J. H. Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. Haematologica 98, 1487–1489 (2013).

13. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).

14. Ionita-Laza, I., McCallum, K., Xu, B. Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. 48, 214–220 (2016).

15. Quang, D., Chen, Y. Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31, 761–763 (2014).

16. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 15, 480 (2014).

17. Huang, Y. F., Gulko, B. Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. 49, 618–624 (2017).

18. Backenroth, D. et al. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation. Am. J. Hum. Genet 102, 920–942 (2017).

19. Lu, Q., Powles, R. L., Wang, Q., He, B. J. Zhao, H. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLOS Genet. 12, e1005947 (2016). 20. Zhou, J. Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods 12, 931–934 (2015). 21. Zou, H. Hastie, T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. 67, 301–320 (2005).

22. Friedman, J., Hastie, T. Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1 (2010).

23. Prentice, R. L. Pyke, R. Logistic disease incidence models and case-control studies. Biometrika 66, 403–411 (1979).

24. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394 (2012).

25. Li, M. J. et al. Predicting regulatory variants with composite statistic. Bioinformatics 32, 2729–2736 (2016). 26. Brown, C. D., Mangravite, L. M. Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLOS Genet. 9, e1003649 (2013).

27. Farh, K. K. H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).

28. Maurano, M. T. et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat. Genet. 47, 1393–1401 (2015).

29. Brown, A. A. et al. Predicting causal variants affecting expression using wholegenome sequencing and RNA-seq from multiple human tissues. Nat. Genet. 49, 1747–1751 (2017).

30. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 (2015).

31. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014).

32. Forrest, M. P. et al. Open chromatin profiling in hipsc-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. Cell Stem Cell 21, 305–318 (2017).

33. Duan, J. et al. A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. Am. J. Hum. Genet. 95, 744–753 (2014).

34. Lee, S., Abecasis, G. R., Boehnke, M. Lin, X. Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 95, 5–23 (2014).

35. He, Z., Xu, B., Lee, S. Ionita-Laza, I. Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. Am. J. Hum. Genet. 101, 340–352 (2017).

36. Voight, B. F. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLOS Genet. 8, e1002793 (2012).

37. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).

38. Liu, D. J. et al. Exome-wide association study of plasma lipids in > 300,000 individuals. Nat. Genet. 49, 1758–1766 (2017).

39. Lu, X. et al. Exome chip meta-analysis identifies novel loci and East Asianspecific coding variants that contribute to lipid levels and coronary artery disease. Nat. Genet. 49, 1722–1730 (2017).

40. Roussos, P. et al. A role for noncoding variation in schizophrenia. Cell Rep. 9, 1417–1429 (2014).

41. Belkin, M., Niyogi, P. Sindhwani, V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7, 2399–2434 (2006).

42. Jiang, Y., He, Y. Zhang, H. Variable selection with prior information for generalized linear models via the prior Lasso method. J. Am. Stat. Assoc. 111, 355–376 (2016).