May 29, 2020

Predicting effects of non-coding variants with a deep learning-based sequence model.

Reference Paper : *Predicting effects of non-coding variants with deep learning-based sequence model* by Zhou and al. (2015)

Vincent Bernaert

Submitted to:
Prof. Jacques Rougemont

Supervised by:
Vojislav Gligorovski

https://github.com/bernaert/BIO-463-Genomics-and-Bioinformatics

1 Introduction and project area

Predicting the functional impact of genetic mutations is key to develop personalized medicine strategies for treating complex diseases with a strong genetic component (i.e. Cancer, Type II Diabetes) [1]. Mutations predisposing an individual to certain diseases can be identified through experimental and observational genome wide association studies (GWAS). However, the exact mechanism by which the mutation causes the disease is often unknown. Moreover, finding causal mutations within the very large number of silent mutations is challenging. This is especially true for single base polymorphisms (SNPs), which are very common single base substitutions ubiquitously distributed throughout the whole genome. It is thought that about half of all SNPs are located in non-coding regions of the genome [1], indirectly interfering with protein expression by modulating regulatory mechanisms such as transcriptional activity or epigenetic marks. To gain insight into the role of these SNPs, many research groups have developed bioinformatics tools aimed to characterize functional SNPs at the genome-scale. Although these efforts had long been limited to the study of SNPs located in coding regions [2], the increasing availability of large experimental datasets like the Encyclopedia of DNA elements (ENCODE, [3]) have allowed researchers to address the problem of non-coding variant prediction.

With large datasets come new computational methods to solve these problems. In particular, supervised machine learning models efficiently extract meaningful information from large datasets containing many different features [4]. This is particularly useful for solving complex classification problems where the final decision depends on many different parameters. Some of these algorithms are used to build models which predict the functional effects of sequence variants, a problem where integrating a wide range of different data types and formats such as contextual chromatin information is crucial. For example, Ritchie et al. [2] used a Random Forest classifier to identify functionally relevant variants in non-coding regions. In the present study, a Deep Learning model based on the convolutional neural network architecture [5] was used to solve the variant prediction problem. The functional 'chromatin features' predicted by the model include transcription factor binding affinity, histone marks and DNase I-hypersensitive sites. The model is trained on sequences annotated for these features and learns associations between them. This results in a model capable of simultaneously predicting the functional effects of non coding variants on 919 different chromatin features.

2 Model and data

2.1 DeepSEA model description

The input data to the DeepSEA is generally a set of 1000-bp sequences taken from anywhere in the genome with coordinates in the GRCh37 reference genome format [6]. Each input 1000-bp sequence is one hot-encoded to represent the categorical variables (bases) as binary vectors. The model itself is a classifier and is composed of three parts. The first part extracts features by successively applying convolutional filters to the input. The second part integrates all of the information extracted in the previous one through a fully connected layer. The final part uses the sigmoid activation function to compute an output probability vector of length 919, which corresponds to the number of chromatin features predicted. To predict the effect of a single base variant (SNP), one can feed the model with the SNP information in the right format containing: the position on the chromosome, the reference allele and the alternative allele. The model considers this input as two 1000-bp sequences each centered around the reference and alternative allele and computes two output probability vectors for each. To determine if a chromatin feature is enriched in the SNP, Log2 fold changes between the reference and alternative output probability vectors are computed.

2.2 Data, training and testing procedure

The data used to train the model is taken from the ENCODE and Roadmap Epigenomic projects. It consists of a set 919 chromatin profiles in narrowPeak (BED4+6) format: 104 histone marks (HM) profiles, 125 Dnase Hypersentitive Sites (DHS) profiles and 690 Transcription factor (TF) profiles. Each file contains called peaks based on pooled and normalized experimental data from a combination of cell type, feature and treatment. HM and TF profiles were obtained from Histone Chip-seq (TF Chip-seq) experiments and DHS from DNase-seq experiments. Quality controls such as the number biological replicates, the number of fragments per replicate, an evaluation of antibody quality and input controls were all performed before data release. The training input for the DeepSEA model is a set of n 1000-bp sequences each paired with a (1x919) vector of boolean values: '1' if the chromatin feature is enriched in the input sequence and '0' otherwise. The weights of this model are calculated by minimizing an objective function (negative log likelihood) using the stochastic gradient descent algorithm. The output of each sequence sample is a (1x919) vector of normalized probabilities for all chromatin features. A higher probability for a given chromatin feature in the output vector indicates that this chromatin feature is more likely to be enriched in the input sequence. As with any binary classifier, the decision to classify a sample as positive or negative for a given feature depends on a predefined threshold.

3 Evaluating the DeepSEA model

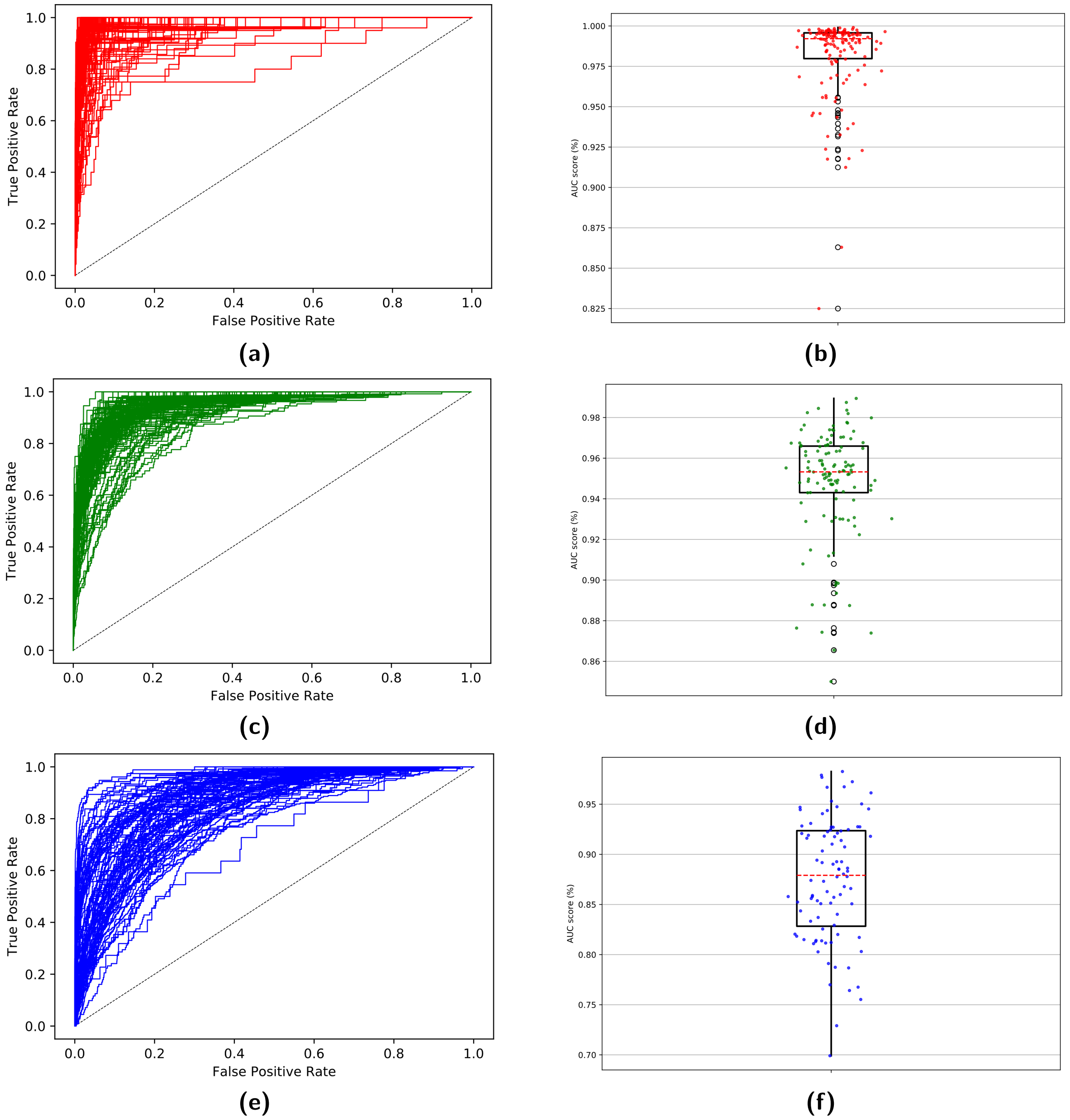
3.1 Methods

The aim of this section is to evaluate the performance of the DeepSEA model. For this task, we first generated a test dataset containing a total of 5000 1000-bp long DNA sequences randomly located on chromosome 9 spanning coordinates 30,000,924 - 38,000,661. Given that chromosome 9 was not used for training the model in the original paper, it should provide an unbiased evaluation of the model's performance.

For each sample in the test dataset, we generated a binary label vector for the 919 chromatin features using the same method as the authors: a feature was labeled as class '1' if at least half of the central 200-bp in the 1000-bp sequence is contained in in the peak region of the corresponding chromatin profile and '0' otherwise. Different metrics can be used to evaluate the performance of deep-learning classifiers. Since we are dealing with a binary classification problem where the two classes are imbalanced (enrichment of a feature is a rare event), a simple measure of accuracy is clearly not sufficient. Instead, Receiver operating characteristic (ROC) curves showing the ratio of true positives (TP) vs. false positives at multiple thresholds are calculated for each feature (Fig. 1). To assess the model performance on each feature, area under the curve (AUC) scores are computed (Fig. 1). A perfect model with an AUC score of 1 classifies all samples correctly.

3.2 Results

The best test performance was attained for transcription factors with an average AUC score of 0.982, followed by DNase sites (0.948) and histone marks (0.875). All 919 features were tested. For comparison, a random classifier has a score of 0.5. The lower performances for Dnase sites and histone marks in particular could be due to the fact that these features span longer regions of the chromosome and therefore the probability of classifying a sample as a false positive is higher. Nonetheless, the overall performance of the DeepSEA model is high and consistent with results from the original paper.



(a) (b)

(c) (d)

(e) (f)

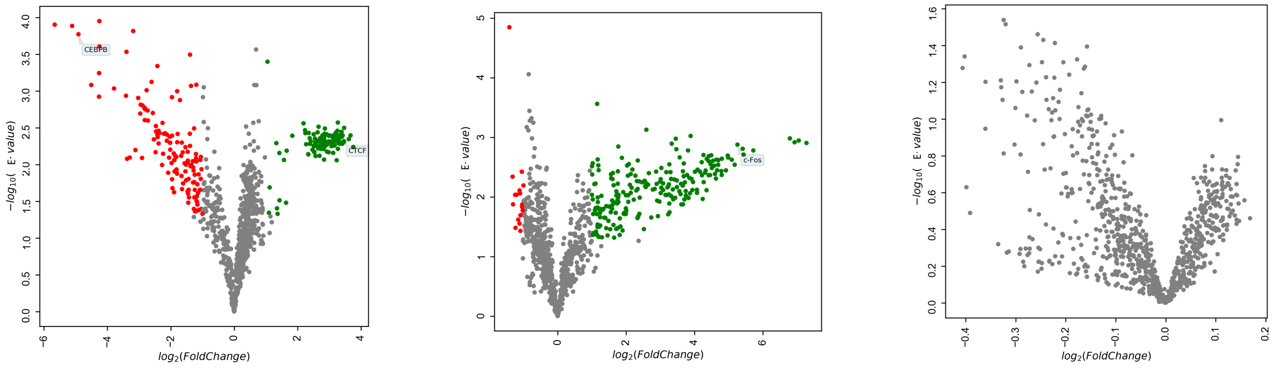
Figure 1: Evaluation of DeepSEA model performance
ROC curves and the corresponding AUC scores. Each curve and point on the boxplot corresponds to a prediction for one of the 919 features on the whole test set (5000 1000-bp sequences spanning coordinates 30,000,924 - 38,000,661 on chromosome 9). (a), (b) Transcription Factors (AUC = 0.982 ± 0.025. (c) , (d) DNase Hypersensitive Sites (AUC = 0.948 ± 0.027). (e), (f) Histone Marks (AUC = 0.875 ± 0.061). Only chromatin features with at least 20 out of 5000 positive samples in the original test set are used. The black dashed line represents the performance of a random classifier. This figure is inspired from Fig 2 of [7].

4 New Analysis 1: Diabetes associated SNPs

4.1 Methods

In this section we focus on predicting variants (SNPs). We first gathered a set of diabetes-associated variants previously selected from the GWAS catalog by Bucher et al. [8]. The original dataset contains the rsIDs of 817 SNPs. Using the Ensembl Effect Predictor tool, we converted SNP rsIDs into genomic coordinates in the standard .vcf format. Finally, we converted these coordinates to the GRCh37/hg19 reference genome as required by the DeepSEA model using the UCSC Genome Browser LiftOver tool. After removing duplicates and variants with unresolved alternative alleles, the dataset contains a total of 638 diabetes associated SNPs.

Here, we are interested in identifying possible relationships between TF binding affinity and histone marks enrichment in the diabetes SNP dataset. For example, a given SNP may increase the binding affinity for a specific TF and we may be interested in knowing if this event is also associated with an increase of certain histone mark near that SNP. For this, the dataset of 638 SNPs is first passed to the DeepSEA model. The output of this procedure is a (636x919) matrix of Log2 fold change values of all chromatin features for each SNP. The effects of individual SNPs on chromatin features can be visualized with volcano plots (Figure 2). For example, Figure 2.a shows a cluster of enrichment for all CTCF features (TF), Figure 2.b shows enrichment of c-Fos (TF) and Figure 2.c shows no significant change for any of the features.



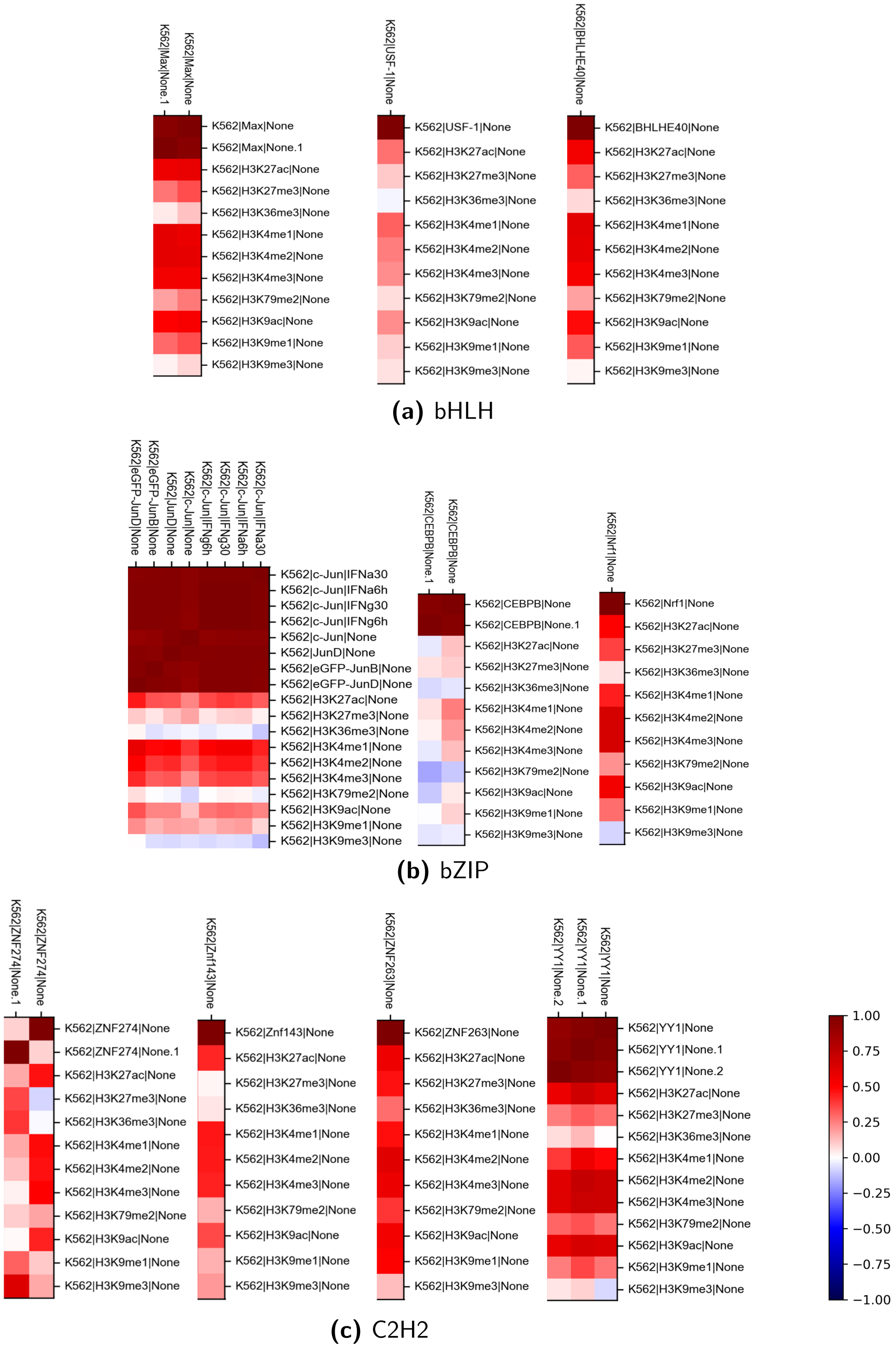
(a) **(b)** **(c)**

Figure 2: Volcano plots of chromatin features enrichment from 3 diabetes associated SNPs
Green: the feature is enriched in the SNP. Red: the feature is depleted in the SNP. (a) SNP on chr11 at location 8243798 (A to G) , (b) SNP on chr21 at location 42583738 (G to A), (c) SNP on chr10 at location 97284081 (C to T). Each point corresponds to one of the 919 chromatin features predicted by the DeepSEA model. On the x-axis : Log 2 fold change of the predicted probability for the alternative allele (SNP) vs. the reference allele (WT). y-axis : E-value, defined as the expected proportion of SNPs with larger predicted effect (from reference allele to alternative allele) for this chromatin feature. E-value is computed based on the empirical distributions of predicted effects for 1000 Genomes SNPs [7]. The higher the -log(E), the better the prediction. Volcano plots were made using the bioinfokit toolkit [9].

To find association patterns between various histone marks and transcription factor binding sites (TFBS), we then calculated pearson correlation plots based on the Log 2 fold changes predicted by the DeepSEA model. Correlation results (Figure 3) are calculated for the K562 cell type and for three different transcription factor families: C2H2, bHLH, bZIP. These three transcription factor families were chosen as they have been previously shown to be highly associated to histone marks *in vitro* [10]. A high positive value indicates that the SNP has the same effect on the transcription factor binding affinity and the histone mark (increase or decrease). Conversly, a low negative value indicates no association between TF binding affinity and level of histone modification.

4.2 Results

For the bHLH family of transcription factors (Figure 3.a), the pattern of histone mark levels seems conserved. H3K27ac, H3K9ac , H3K4me1, H3K4me2 and H3K4me3 show the highest levels of correlation across all three transcription factors. All 5 histone marks have been previously associated with an upregulation of gene transcription [11]. H3K36me3 is either weakly correlated or even negatively correlated to USF-1. The bZIP transcription factors (Figure 3.b) also share similar patterns of histone mark activation with some differences in the correlation values. H3K4me1, H3K4me2, H3K4me3 are all positively correlated with an increase transcription factor affinity. H3K9me3 and H3K36me3 are both negatively correlated or not correlated with transcription factor binding affinity. However, the CEBP subtype shows weaker correlation values across all histone marks, indicating that the transcription factor binding site may be depleted in the alternative allele. We obtain similar results for the C2H2 family with mostly conserved association patterns (Figure 3.c). As expected, the strongest associations found for all three transcription factor families are with enhancer histone marks H3K4me1, H3K4me2 and H3K4me3. These results indicate that the global patterns of histone mark associations are conserved within each transcription factor family and that the strongest associations (H3K4me1, H3K4me2 and H3K4me3) are conserved for all three families.



(a) bHLH

(b) bZIP

(c) C2H2

Figure 3: Heatmap of correlation results between histone marks and transcription factor Log2 fold changes for all 638 SNPs. Predicted chromatin features are defined by a combination of cell type—feature—treatment. (a) Transcription factors of the basic helix-loop-helix (bHLH) family, (b) Transcription factors of the Basic Leucine Zipper Domain (bZIP) family, (c) Transcription factors of the C2H2 zinc finger family. Correlation plots were made using the bioinfokit toolkit [9].

5 New Analysis 2: Predicting transcription factor binding motifs

5.1 Methods

In this section, we extend the analysis to the prediction of transcription factor binding motifs (TFBM). We begin by identifying an SNP which shows an increase in affinity for the TF we are trying to predict. Indeed, an SNP could create a new TFBS absent in the reference allele. We then perform random *in silico* point mutagenesis on the bases flanking the SNP and check whether mutating each base leads to an increase or a decrease in TF binding affinity. For this task, the Sequence Profiler tool from the DeepSEA model which performs computational mutagenesis and evaluates the effects of mutating every base of the input sequence on chromatin feature prediction was used. The effect on the chromatin feature is measured as the Log2 fold change in enrichment probability between the reference and alternative allele.

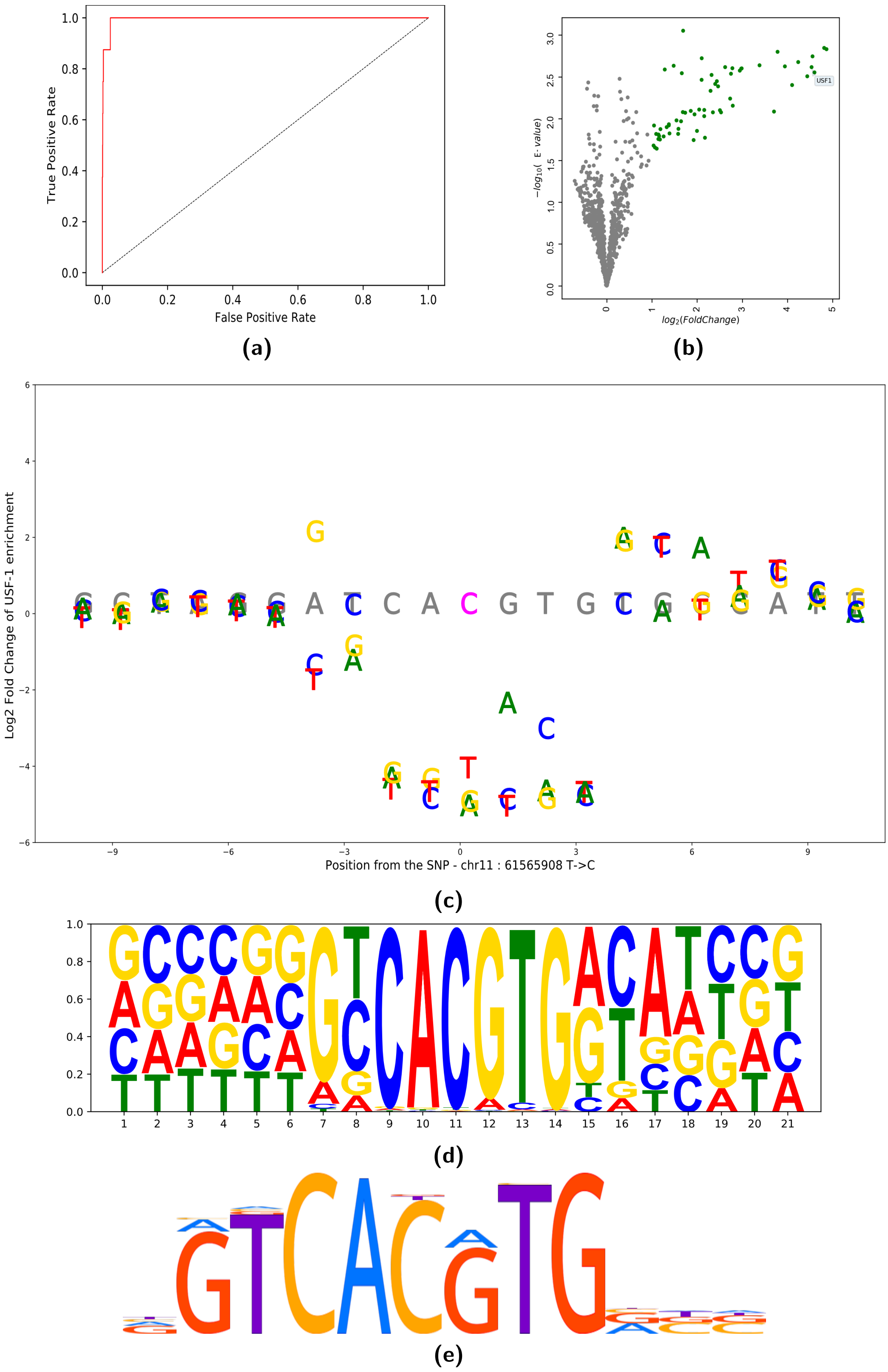
We chose to predict the transcription factor binding motif (TFBM) for USF-1 of the bHLH family in HepG2 cells, for which the DeepSEA model shows high performance (Figure 4.a). In diabetes, the USF-1 in HepG2 feature is enriched in the SNP at position 61565908 of chromosome 11 (T to C, Figure 4.b), which was used as the baseline sequence to perform the analysis. The effect of mutating each of the bases neighboring the SNP on TF binding affinity are shown in the 'Floating Bases' graph (Figure 4.c). For each position, three Log2 fold change values are calculated. For example, if the reference base in the original sequence was 'G', Log2 fold change values for replacing the reference base with alternatively a 'T', 'A' and a 'C' are computed. Using this data, we constructed a Position Probability Matrix (PPM) (Figure 4.d) to evaluate the relative importance of each base at every position surrounding the SNP on USF-1 binding affinity. Given a set of 4 Log2 fold change values, we used a Softmax transformation to convert these values into normalized probabilities. For a given position, the relative contribution of each base in terms of probability is given by :

$$softmax(x)_i = \frac{exp(x_i)}{\sum_j exp(x_j))}$$

where x is the Log2 fold change value for each base with respect to the reference base (note that the reference base always has a null Log2 fold change value).

5.2 Results

The predicted motif (Figure 4d) is very similar to the available motif for USF-1 (Figure 4e), suggesting that the method can efficiently predict transcription factor binding motifs. From these results, the consensus sequence for USF-1 seems to be 'CACGTG', which explains why the mutation from T to C at position 61565908 of chromosome 11 initially showed an increase for USF-1 binding affinity. Overall, these results suggest that the DeepSEA model can be extended to the prediction of transcription factor binding motifs.



(a)

(b)

(c)

(d)

(e)

Figure 4: Predicting USF-1 binding motif from the diabetes associated SNP chr11-61565908 (a) ROC curve for performance of the DeepSEA model on feature USF-1 in HepG2 cells based on the test data from Section 3 (with 8 positive samples). (b) The chromatin feature USF-1 in HepG2 is enriched in the SNP. (c) Log2 fold changes for each base mutation with respect to the reference base (in grey). Magenta: base corresponding to the SNP. (d) Predicted USF-1 binding motif. (e) USF-1 binding motif obtained from the HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) database [12].

6 Conclusion and Discussion

In their original study, Zhou et al. developed a new computational method based on deep learning to predict the effects of sequence variants. In this report, we initially tested the claims made by the authors regarding the model's performance. Our results confirm that the DeepSEA model has good predictive power in general. In a secondary analysis, we gathered a set of SNPs linked to diabetes and studied relationships between histone marks and transcription factor binding affinities. One drawback of our analysis is that it lacks reference data. Although it may provide hints for elucidating the effects of SNPs in diabetes, it is not possible to say if the observed associations between histone marks and TF binding affinity are specific to diabetes. Furthermore, correlation alone is a rather weak metric to measure relationships between chromatin features. For example, it is impossible to determine if TF binding affinity is caused by the histone marks or vice-versa. To improve the analysis, it would be interesting to add a random control SNP dataset for comparison as well as experimental data to confirm the results. In the final analysis, we proposed a method based on DeepSEA which can efficiently predict TF binding motifs. The fact that we were able to recover the correct motif for a TF confirms that DeepSEA is indeed sensitive to single-base mutations. This method could be applied for the prediction of any TFBM, provided that the model has a high AUC score for that TF.

References

[1] B. S. Shastry, "SNP alleles in human disease and evolution," 2002.

[2] G. R. Ritchie, I. Dunham, E. Zeggini, and P. Flicek, "Functional annotation of noncoding sequence variants," *Nature Methods*, vol. 11, pp. 294–296, 2 2014.

[3] I. Dunham, A. Kundaje, and Aldred, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, 9 2012.

[4] T. Hillel, M. Bierlaire, and Y. Jin, "A systematic review of machine learning methodologies for modelling passenger mode choice," tech. rep., 2019.

[5] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks,"

[6] "GRCh37 - hg19 - Genome - Assembly - NCBI."

[7] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Methods*, vol. 12, pp. 931–934, 9 2015.

[8] S. Kumar, G. Ambrosini, and P. Bucher, "SNP2TFBS-a database of regulatory SNPs affecting predicted transcription factor binding site affinity," *Nucleic Acids Research*, vol. 45, pp. 139–144, 2016.

[9] R. Bedre, "reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit," 5 2020.

[10] B. Xin and R. Rohs, "Relationship between histone modifications and transcription factor binding is protein family specific," *Genome Research*, vol. 28, pp. 321–333, 3 2018.

[11] A. Sharifi-Zarchi, D. Gerovska, K. Adachi, M. Totonchi, H. Pezeshk, R. J. Taft, H. R. Schöler, H. Chitsaz, M. Sadeghi, H. Baharvand, and M. J. Araúzo-Bravo, "DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism," *BMC Genomics*, vol. 18, 12 2017.

[12] "NANOG_MOUSE.H11MO.1.A motif - HOmo sapiens COmprehensive MOdel COllection."