**Predicting effects of noncoding variants with deep learning based sequence modelb**

**Abstract e**

Identifying functional effects of noncoding variants is a major challenge in human genetics. To **predict the noncoding-variant effects *de novo* from sequence**, we developed a deep learning–based algorithmic framework, **DeepSEA** (<http://deepsea.princeton.edu/>), that directly learns a regulatory sequence code from large-scale chromatin-profiling data, enabling **prediction of chromatin effects of sequence alterations** with single-nucleotide sensitivity. We further used this capability to improve prioritization of functional variants including expression quantitative trait loci (eQTLs) and disease-associated variants.

**Main**

* The overarching goal of the study is to predict the function of non-coding variants by learning from regulatory sequence information.
* In particular, the method aims to predict with single nucleotide sensitivity (ex: A 🡪 T in a given regulatory DNA sequence) the effect of noncoding variants on **transcription factor binding (TF-binding), DNA accessibility, and histone marks of sequences**.
* The method is based on Deep Learning (DeepSEA). DeepSEA learns to predict chromatin effects of noncoding variants, including **DNase I—hypersensitive sites (DHSs), histone marks, and TF binding.** These effects are linked. For example, DHS regions are more accessible by TF because in these regions, DNA has lost its condensed structure, making it more accessible.
* **Three major features:**

1. integrating sequence information from a wide sequence context (the DNA sequences processed by the model are not limited to the TF binding site).

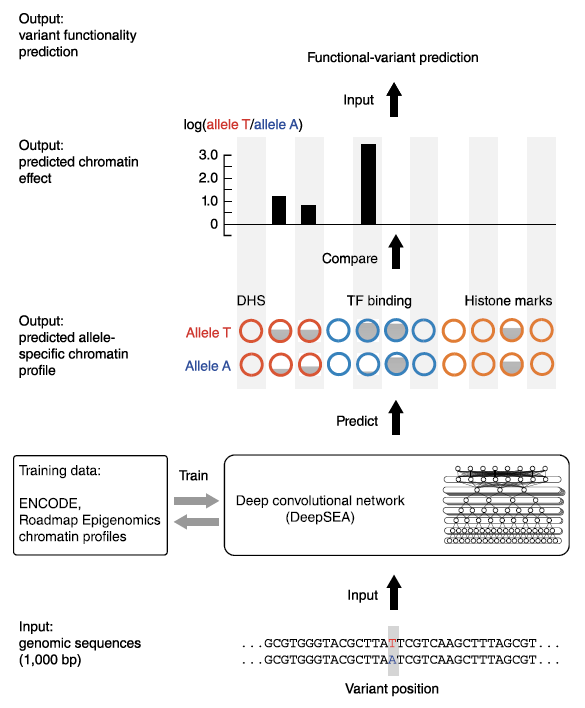
* Considering sequence context is important to determine regulator properties of the variants. To achieve this, increasing the context size to 1 kbp. Substantially increases performance (accuracy of prediction I assume)

1. learning sequence code at multiple spatial scales with a hierarchical structure.

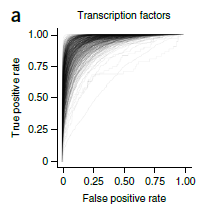
* Heterogeneity (spatial location in genome) of the ‘’learned’’ sequences

1. multitask joint learning of diverse chromatin factors sharing predictive features.

* the model learns to predict across all chromatin predictors. This improves computational efficiency and performance because prediction accuracy is shared across a wide range of feature profiles (TF, DHS, histone marks).
* Example: a given feature effective for predicting binding of a specific TF can be also used by another predictor.
* **Pipeline**

Typical Workflow for predicting TF binding

* Train the model on a dataset with DNA sequences different cell types as input data and binding / or not to a specific TF as the target.
* After training, the model as learned predictors (parameters) for binding to the specific TF.
* Input genomic sequences to the mode (with variants)
* Identify effect of nucleotide alterations on TF binding.
* Plot Receiver Operating Characteristic Curves (**ROC**) for each TF (figure to reproduce)

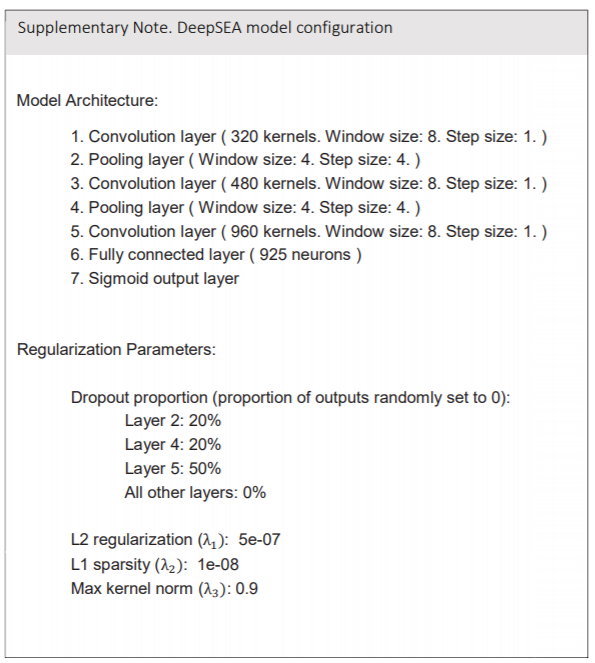


* Guessing the class discrimination threshold is constant and equal to 0.5 here …

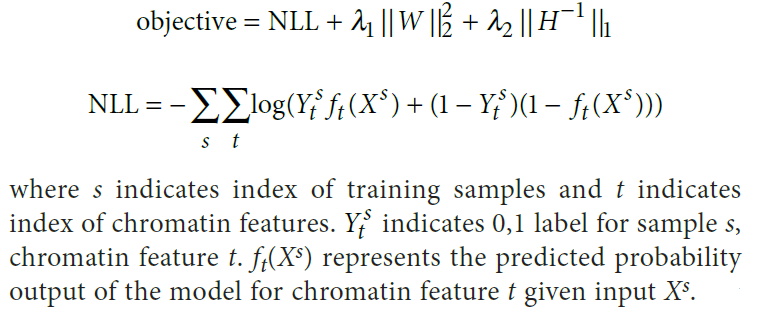
**Online Tools:**

**Model Description**

* Deep Convolutional Neural Network
* Operations: convolution, max pooling, fully connected, sigmoid
* Activation : ReLU
* 3 conv layers (320, 480, 960 kernels) each followed by max pooling layers
* A fully connected layer on top of third convolutional layer with matrix W
* A sigmoid layer on top of the last FCL : makes predictions for 919 features and scales 0-1 to return probabilities for each one (this is where shared predictive sequence features takes place)



**Training**

* The model is a superposition of 925 binary classifiers (0/1) which share the predictive weights
* In its final output layer, the network applies 925 Sigmoids each outputting between 0 and 1
* Minimization of the sum of negative log likelihood (NLL) + regularization is used as the ‘’master’’ objective function for the whole model
* Standard backprop. + SGD with momentum to optimize the predictive weights (pretty standard).

**Input Data for DeepSEA**

Training Data

1. Split genome from a cell type into 200-bp bins (say we have K bins in the genome)
2. For each of the K bins, 919 labels (1 label per chromatin feature)

* **1** : more than ½ of the 200-bp bin is in the peak region
* **0** : otherwise

🡪this is all obtained from Chip-seq experiments

1. Keep the 200-bp bins with at least 1 TF binding event (at least one of the 919 labels is labeled as 1) 🡪 521,636,200 bp of sequence (17% of genome)
2. Each training sample consists of a 1,000 bp sequence (the 200-bp bins kept from before flanked on both sides by two 400-bp regions to provide contextual information)
3. Each training sample is paired with a 919 label vector (0/1) for chromatin features



1. The model processes the 1000 bp sequence as a 1000 x 4 binary one hot encoded matrix with columns corresponding to A, G, C, T.

Data Types



* vcf, fasta, bed
* **vcf** used for predicting non-coding variants m
* to predict feature probabilities, **fasta**
* for specifying sequences from the human reference genome (GRCh37/hg19), **bed** (the bed input should specify regions of 1000-bp length

€awàles

**©☹**

**Glossary:**

* **AUC** (Area Under Curve) : a metric used in classification analysis to determine classification accuracy of a certain predictor (example: AUC for TF binding sites will indicate how well the model is able to predict TF binding sites from DNA sequence variants). (need check more in online tools part as they are many different subtypes of AUC curves).
* **QTL** : Quantitative Trait Loci : a region in the DNA, which is associated with a particular phenotypic trait, which varies in degree. eQTL—expression quantitative trait loci : genomic loci which explain variation in expression levels of mRNAs
* Allelic Imbalance : when one allele is observed in DNA-seq data significantly more often (more reads of the same allele) than other allele at a heterozygous site for single cell-type sample. Example : In a DNAase-seq assay, if one allele is found more often, this indicates that that allele may be more sensitive to DNase I.
* In Silico saturated mutagenesis: computational approach to scan along all potential single nucleotide substitutions and predicts the chromatin effects of each. This method identifies which sequences features are most informative for a specific effect prediction.
* Feature : motifs in the DNA sequence (e.g. chromatin annotations) used by the model to predict effects on chromatin.
* Effects : What is predicted by the model: TF binding sites, DHS accessible regions, regions targeted for histone marks.
* Predictor : the parameters learned by the model to predict. They can be optimized for binding of a specific TF, or of a physically interacting one for example, depending on the training data.
* SNP : Single Nucleotide Polymorphism : a single-nucleotide variant with functional effects.
* Non-coding variants : variants in on-coding DNA sequence (introns) associated with functional effects, but not directly with protein (ex: a mutation in a gene regulatory sequence).
* DNase I-hypersensitive sites (DHSs) : regions of chromatin that are sensitive to cleavage by the DNA I enzyme. In these regions of the genome, chromatin as lost its condensed structure, exposing the DNA and making it more accessible.
* Chromatin : the mass of genetic material composed of DNA and proteins that condense to form the chromosome. (DNA packaging system which can be more or less accessible for epigenetic modifications)
* Epigenetics : modification of gene expression through DNA marks rathe than alteration of the code itself. For example, DeepSea can be used to predict epigenetic effects.
* Receiver Operating Characteristic Curve (ROC) : graphical plot illustrating the diagnostic ability of a binary classifier system as its discrimination threshold is varied (<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>)

**Tools**

-> classic pandas with dataframes

-> python packages such as pybedtools

<http://daler.github.io/pybedtools>

We support only GRCh37/hg19 genome coordinates. You can use LiftOver to convert your coordinates to the correct version.

-> R using R vanilla

<https://charlesjb.github.io/How_to_import_narrowPeak/>

* Genomes database : <https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

Exploratory Data Analysis

* include histograms of Chip-seq peaks for various chromatin features profiles from each type (histone mark, transcription factor, DNase hypersensitive sites)

TASK 1: Evaluate the model by reproducing ROC curves

* Include the curves for ALL Transcription factors, ALL DNase Hypersensitive sites, and ALL Histone Marks
* Include a boxplot for AUCs

TASK 2: Identify a TF which is associated with a clinical phenotype (mutation). Check if the model can predict the change in TF binding affinity and investigate which types of mutations are associated with the phenotype. Basically, check if the model is capable of predicting TF binding changes

Idea :

* Gather all SNPs relevant for a phenotypical trait (ex: cancer aggressiveness) from tools like this:

Liftover tool for converting from 38 to 37

<https://genome.ucsc.edu/cgi-bin/hgLiftOver>

tools to get SNPs

<https://www.ebi.ac.uk/gwas/search?query=rs35148638>

or

<http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=9:9524-10524;v=rs1563683293;vdb=variation;vf=579642862>

* Rank all TFs based on increase of binding affinity (gain of function), or decrease
* Based on these scores, identify TFs which are implicated in disease and check if this makes sense

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501730/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3638136/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6464064/>

‘Case Studies’

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6906655/>

tool which computes the PWM given a TF

<https://hocomoco11.autosome.ru/motif/NANOG_MOUSE.H11MO.1.A>

tools which scans genome and finds PWM hits for TFS

[https://ccg.epfl.ch//pwmtools/pwmscan.php](https://ccg.epfl.ch/pwmtools/pwmscan.php)

scoring and ranking TFs

<https://www.nature.com/articles/s41467-018-04406-2>

TASK 3: Choose a TF for which the model shows high binding site predictability. Perform computational mutational scan to find the consensus sequence of the TF. Compare results with experimentally available determined motif if available.

<https://www.nature.com/articles/ng.2416.pdf>

<https://www.biorxiv.org/content/10.1101/581306v1.full>

Cis-BP database

* Myc max

Relationships between features :

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5848611/>

<https://www.pnas.org/content/pnas/116/9/3668.full.pdf>