Predicting the Structures and Mutations of DNA and RNA using Deep Learning

[1]Dr.A.Sharada, [2]M.Sreevidya, [3]D.Vaishnavi,
[1]Professor, CSE, [2]II year B.Tech, [3]II year B.Tech
G.Narayanamma Institute of Technology and Science,Shaikpet, Hyderabad

ABSTRACT-Technological advances in genomics have led to an explosion of molecular and cellular profiling data from large numbers of samples. This rapid increase in biological data dimension and acquisition rate is challenging traditional analysis strategies. Modern machine learning methods, such as deep learning, promise to leverage very large data sets for finding hidden structure within them, their analysis and for making accurate predictions.

In this review, we discuss applications of this new breed of analysis approaches in regulatory genomics and cellular imaging. We provide background of what deep learning is, and the settings in which it can be successfully applied to derive biological aspects. In addition to presenting specific applications we also highlighted possible pitfalls and limitations to guide computational biologists when and how to make the most of its use of this new technology.

KEYWORDS: *genomics, proteomics, metabolomics, supervised learning, deep learning, deep bind, mutation maps, in vitro , in vivo, convolution neural networks.*

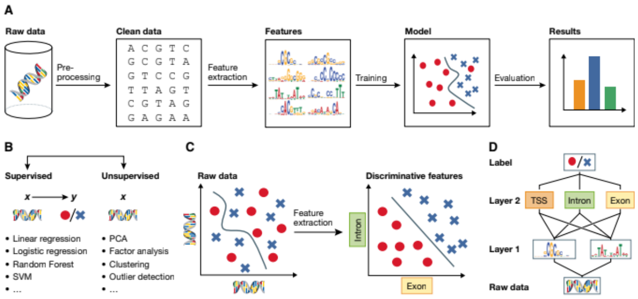
1.INTRODUCTION

In computational biology, the appeal is the ability to derive predictive models without a need for strong assumptions about underlying mechanisms, which are frequently unknown or insufficiently defined. As a case in point, the most accurate prediction of gene expression levels is currently made from a broad set of epigenetic features using sparse linear models or random forests how the selected features determine the transcript levels remains an active research topic. Predictions in genomics, proteomics**,** metabolomics or sensitivity to compounds all rely on machine learning approaches as a key ingredient. Most of these applications can be described within the canonical machine learning workflow, which involves four steps: data cleaning and pre-processing, feature extraction, model fitting and evaluation . It is customary to denote one data sample, including all co-variates and features as input x (usually a vector of numbers),and label it with its response variable or output value y (usually a single number) when available .A supervised machine learning model aims to learn a function f(x) = y from a list of training pairs (x1,y1), (x2,y2), ... for which data are recorded . One typical application in biology is to predict the viability of a cancer cell line when exposed to a chosen drug.

Knowing the sequence specificities of DNA- and RNA-binding proteins is essential for developing models of the regulatory processes in biological systems and for identifying causal disease variants. Here we show that sequence specificities can be ascertained from experimental data with ‘deep learning’ techniques, which offer a scalable, flexible and unified computational approach for pattern discovery. Using a diverse array of experimental data and evaluation metrics , we find that deep learning outperforms other state-of-the-art methods, even when training on in vitro data and testing on in vivo data. We call this approach Deep Bind and have built a stand-alone software tool that is fully automatic and handles millions of sequences per experiment. Specificities determined by Deep Bind are readily visualized as a weighted ensemble of position weight matrices or as a ‘mutation map’ that indicate show variations affect binding within a specific sequence.

DNA- and RNA-binding proteins play a central role in gene regulation, including transcription and alternative splicing. The sequence specificities of a protein are most commonly characterized using position weight matrices1 (PWMs), which are easy to interpret and can be scanned over a genomic sequence to detect potential binding sites.

Deep Bind addresses the above challenges. (i) it can learn from millions of sequences through parallel implementation on a graphics processing unit (GPU); (ii) It can be applied to both microarray and sequencing data; (iii) it can tolerate a moderate degree of noise and mislabeled training data; (iv) it generalizes well across technologies, even without correcting for technology-specific biases; and(v) it can train predictive models fully automatically, alleviating the need for careful and time-consuming hand-tuning. Importantly, a trained model can be applied and visualized in ways that are familiar to users of PWMs.



2.TWO EXPLORED DOWNSTREAM APPLICATIONS

1. Analyzing disease-associated genetic variants that can affect transcription factor binding and gene expression and,

2. Uncovering the regulatory role of RNA binding proteins (RBPs) in alternative slicing.

2.1 TRAINING DEEP BIND AND SCORING SQUENCES

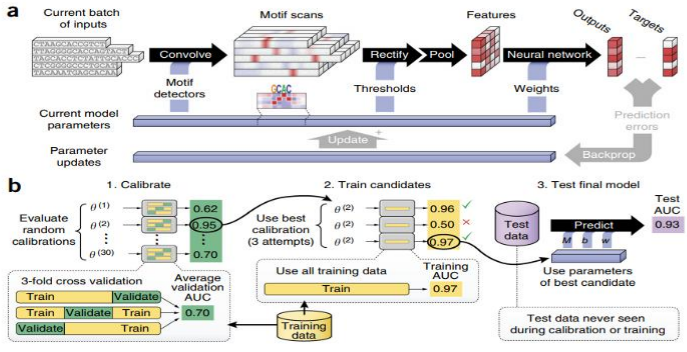
For training, Deep Bind uses a set of sequences and, for each sequence, an experimentally determined binding score. Sequences can have varying lengths , and binding scores can be real-valued measurements or binary class labels.

For sequences, DeepBind computes a binding score f (s) using four stages:

1. $f(s)=net_W(pool(rect_b(conv_M(s))))$.The convolution stage (convM) scans a set of motif detectors with parameters M across the sequence. Motif detector Mk is a 4 × m matrix, much like a PWM of length m but without requiring coefficients to be probabilities or log odds ratios. The rectification stage isolates positions with a good pattern match by shifting the response of detector Mk by clamping all negative values to zero.

2. The pooling stage computes the maximum and average of each motif detector's rectified response across the sequence; maximizing helps to identify the presence of longer motifs, whereas averaging helps to identify cumulative effects of short motifs, and the contribution of each is determined automatically by learning.

3. These values are fed into a nonlinear neural network with weights W, which combines the responses to produce a score. Ascertaining DNA sequence specificities to evaluate Deep Bind's ability to characterize DNA-binding protein specificity, we used PBM data from the revised DREAM5 TF-DNA Motif Recognition Challenge.



2.2 DETAILS OF INNER WORKINGS OF DEEP BIND AND ITS TRAINING PROCEDURE

(a) Five independent sequences being processed in parallel by a single Deep Bind model. The convolve, rectify, pool and neural network stages predict a separate score for each sequence using the current model parameters .

During the training phase, the back prop and update stages simultaneously update all motifs, thresholds and network weights of the model to improve prediction accuracy.

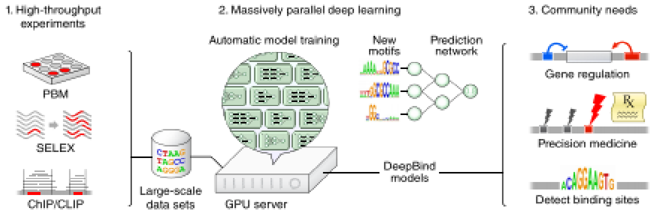
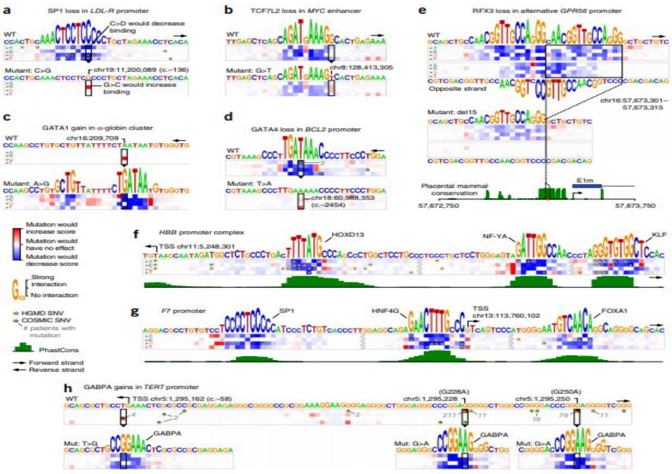


Figure 1 DeepBind's input data, training procedure and applications. 1. The sequence specificities of DNA- and RNA-binding proteins can now be measured by several types of high-throughput assay, including PBM, SELEX, and ChIP- and CLIP-seq techniques. 2. DeepBind captures these binding specificities from raw sequence data by jointly discovering new sequence motifs along with rules for combining them into a predictive binding score. Graphics processing units (GPUs) are used to automatically train high-quality models, with expert tuning allowed but not required. 3. The resulting DeepBind models can then be used to identify binding sites in test sequences and to score the effects of novel mutations.

(b) The calibration, training and testing procedure used throughout .

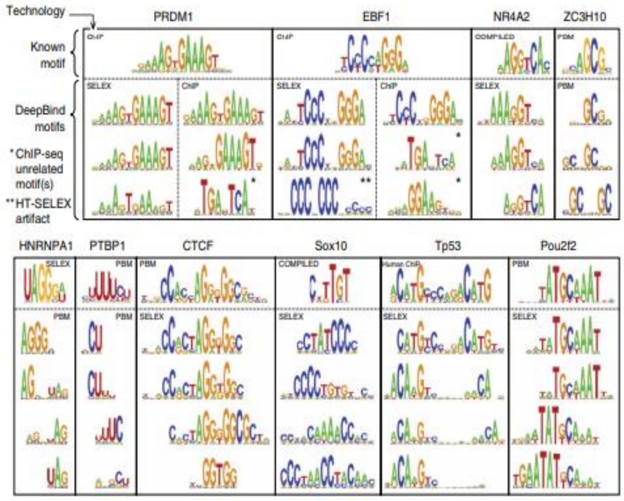
2.3 ANALYSIS OF POTENTIALLY DISEASE-CAUSING GENOMIC VARIANTS

Deep Bind mutation maps were used to understand disease-causing SNVs associated with transcription factor binding. (a) A disrupted SP1 binding site in the LDL-R promoter that leads to familial hypercholesterolemia. (b) A cancer risk variant in a MYC enhancer weakens a TCF7L2 binding site. (c) A gained GATA1 binding site that disrupts the original globin cluster promoters. (d) A lost GATA4 binding site in the BCL-2 promoter, potentially playing a role in ovarian granulosa cell tumors. (e) Loss of two potential RFX3 binding sites leads to abnormal cortical development. HGMD SNVs disrupt several transcription factor binding sites in the promoters of HBB and F7, potentially leading to β-thalassemia and hemophilia, respectively. (h) Gained GABP-α binding sites in the TERT promoter, which are linked to several types of aggressive cancer. WT, wild type.



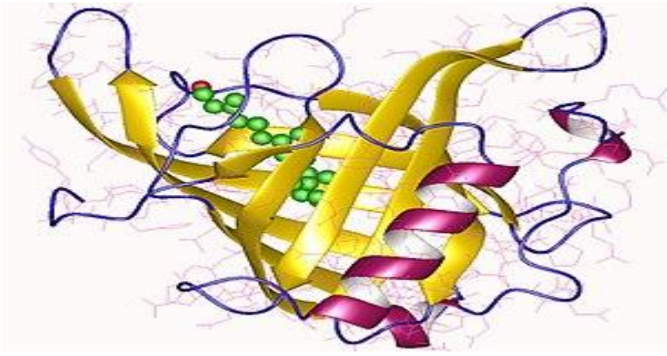
2.4 IDENTIFYING AND VISUALIZING DAMAGING GENETIC VARIANTS

Genetic variants that create or abrogate binding sites can alter gene expression patterns and potentially lead to diseases.



A promising direction in precision medicine is to use binding models to identify, group and visualize variants that potentially change protein binding. To explore the effects of genetic variations using Deep Bind, we developed a visualization called a ‘mutation map’, which illustrates the effect that every possible point mutation in a sequence may have on binding affinity. A mutation map conveys two types of information. First, for a given sequence, the mutation map shows how important each base is for the Deep Bind analysis by the height of the base letter. Second, the mutation map includes a heat map of size 4 by n, where n is the sequence length, indicating how much each possible mutation will increase or decrease the binding score.

3. DEEP BIND MODELS ARE USED TO DESCRIBE THE REGULATION MECHANISM FOR DIFFERENT RBPs: **Retinol-binding proteins** (RBP) are a family of proteins with diverse functions. They are carrier proteins that bind retinol. Retinol and retinoic acid play crucial roles in the modulation of gene expression and overall development of an embryo. In vitro models are consistent with known splicing patterns. RBPs play a crucial role in regulating splicing , having an impact on a wide variety of developmental stages such as stem cell differentiation and tissue development.



4.PRINCIPLES OF USING NEUTRAL NETWORKS FOR PREDICTING MOLECULAR TRAITS FROM DNA SEQUENCE:

(A) DNA sequence and the molecular response variable along the genome for three individuals. Conventional approaches in regulatory genomics consider variations between individuals, whereas deep learning allows exploiting intra-individual variations by tiling the genome into sequence DNA windows centered on individual traits, resulting in large training data sets from a single sample.

(B) One-dimensional convolutional neural network for predicting a molecular trait from the raw DNA sequence in a window.

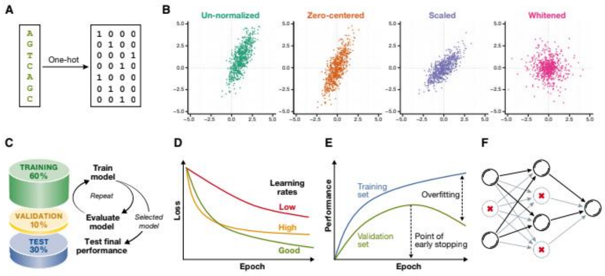
Filters of the first convolutional layer (example shown on the edge) scan for motifs in the input sequence. Subsequent pooling reduces the input dimension, and

additional convolutional layers can model interactions between motifs in the previous layer.

(C) Response variable predicted by the neural network shown in(B) for a wild-type and mutant sequence is used as input to an additional neural network that predicts a variant score and allows to discriminate normal from deleterious variants.

(D) Visualization of a convolutional filter by aligning genetic sequences that maximally activate the filter and creating a sequence motif.

(E) Mutation map of a sequence window. Rows correspond to the four possible base pair substitutions, columns to sequence positions. The predicted impact of any sequence change is colour -coded.

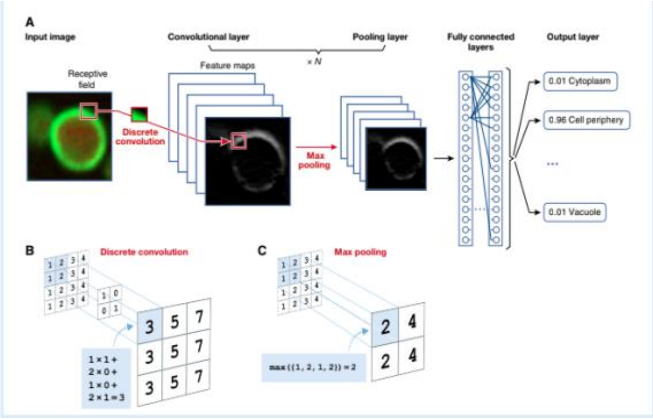


Letters on top denote the wild-type sequence with the height of each nucleotide denoting the maximum effect across mutations. Convolutional network architectures are considered to predict specificities of DNA- and RNA-binding proteins.

Their Deep Bind model outperformed existing methods, was able to recover known and novel sequence motifs, and could quantify the effect of sequence alterations and identify functional SNVs(single nucleotide variants). A key innovation that enabled training the model directly on the raw.

DNA sequence was the application of a one-dimensional convolutional layer. Intuitively, the neurons in the convolutional layer scan for motif sequences and combinations thereof, similar to conventional position weight matrices .

Deep learning for biological image analysis . Historically, perhaps the most important successes of deep neural networks have been in image analysis. Deep architectures trained on millions of photographs can famously detect objects in pictures better than humans do.



All current state-of-the-art models in image classification, object detection, image retrieval and semantic segmentation make use of neural networks.

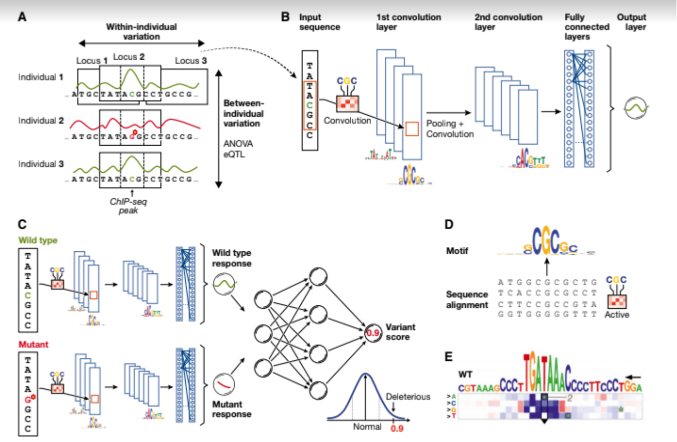
The convolutional neural network is the most common network architecture for image analysis. Briefly, a CNN performs pattern matching (convolution) and aggregation (pooling) operations.

At a pixel level, the convolution operation scans the image with a given pattern and calculates the strength of the match for every position. Pooling determines the presence of the pattern in a region, for example by calculating the maximum pattern match in smaller patches (max-pooling), thereby aggregating region information into a single number.

The successive application of convolution and pooling operations is at the core of most network architectures used in image analysis.

4.1 ANALYSIS OF WHOLE CELLS,CELL POPULATIONS AND TISSUES:

In many cases, pixel-level predictions are not required. For example, directly classified colon histopathology images into cancerous and non-cancerous, finding that supervised feature learning with deep networks was superior to using handcrafted features. This approach allowed classifying entire images without performing segmentation as a pre-processing step. CNNs have even been applied to count bacterial colonies in agar plates also.



5 CONCLUSION:

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep convolutional networks have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech. Research is on to develop algorithms to suit drug designing industry .These networks integrated the techniques of bioinformatics and computational biology. These also helped in discovering and analyzing the intrinsic structures of DNA and RNA and hence enhancing the field of computational biology.

REFERENCES:

[1] Stormo, G. DNA binding sites: representation and discovery. Bioinformatics.

[2]Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J,Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Josofowicz R, KaiserL, Kudlur M, Levenberg J (2016) Tensor Flow.

[3] Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA- binding proteins by deep learning.

[4] Siggers, T. & Gordân, R. Protein-DNA binding: complexities and multi-protein codes.

[5] Mukherjee, S. et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.