Deep Learning for Computational Biology

CS 260

Welcome to CS 260

• Coordinator: Stefano Lonardi
  – Office: WCH room 325
  – Phone: (951) 827-2203
  – Email: stelo@cs.ucr.edu

• Lectures: MWF, 3:10-4pm, WCH 139

• Office hours: by appointment

• http://www.cs.ucr.edu/~stelo/
  (click on “Teaching”, then CS 260 Winter 18)
Course organization

• The course will be structured as a “journal club”, where students alternate presenting papers and highlight and discuss possible new line of research
• Two 20 minutes presentation each lecture
• Some time for discussion
• Interactive!

Course organization

• The final grade will be based on paper presentations and participation in the class
• While the presenter is responsible to describe the paper in detail, it would be beneficial for the others to review the paper in advance to be able to participate in the discussion
• I will have a selection of papers, but students are welcome to propose other papers to present with my OK
Structure of the presentation

• Introduction (background, motivations)
• Statement of the problem
• Proposed solution
• Experimental results (comparison with previous approaches, if applicable)
• Discussion (your review)
  • (please do not waste time with “outline of my talk”)

Introduce yourself

• Name
• Department
• Graduate/undergraduate
• Year at UCR
• Research interests
Proteins
3D structure prediction
secondary structure prediction
docking
Proteins

- A protein is a chain of molecules, called amino acids.
- Every amino acid has a central carbon atom, known as alpha carbon ($C_\alpha$), an amino group ($NH_2$), a carboxyl group (COOH) and a side chain.
- The side chain is what distinguishes one amino acid from another.

Amino acids
Proteins

• Amino acids are linked by *peptide bonds* between the *carboxyl group* and the *amino group*

• Almost all organisms share the same 20 amino acids

• A typical protein is composed by 300 amino acids, but there are proteins with as few as 100 or with as many as 5,000 amino acids

Proteins

• *Primary* structure: the linear sequence of amino acids, ordered from the N-terminal (amino group) to C-terminal (carboxy group)

• *Secondary* structure: α-helices and β-sheets

• *Tertiary* structure: the 3D conformation (*folding*) in space
- Most of the backbone is rigid
- The chemistry of a protein forces most of the backbone to remain planar
- The chemical bonds to the alpha carbons can rotate
- The angle of rotation for each alpha carbon bonds are called $\phi$ and $\psi$
- Phi and psi are the degree of freedom of the protein

**Alpha helix**

- Exactly 3.6 residues per turn
- Hydrogen bonds
- Two types
  - Right-handed
  - Left-handed
Beta sheet

- Regions of extended (nearly linear) backbone conformation with $\phi \approx 135$ and $\psi \approx 135$
- Hydrogen bonds
- Two types
  - Parallel
  - Anti-parallel

Protein structure

- The *function* of a protein is determined by its tertiary structure
- Structure is much more conserved than sequence
- Predicting the folding from the primary sequence is very hard (see CASP competition)
- *Binding*: the interaction between two or more proteins (or protein-DNA) which have a “compatible” 3D structure (*docking*)
(A) The folding of the polypeptide chain typically creates a crevice or cavity on the protein surface. This crevice contains a set of amino acid side chains disposed in such a way that they can make bonds only with certain ligands. (B) Close-up view of an actual binding site showing the hydrogen bonds and ionic interactions formed between a protein and its ligand (in this example, cyclic AMP is the bound ligand).

**Intrinsically disordered proteins**

- Proteins lacks a fixed or ordered three-dimensional structure are called *intrinsically disordered protein*
- IDPs cover a spectrum of states from fully unstructured to partially structured
- Long (>30 residue) disordered segments occur in a third of eukaryotic proteins
- Many IDPs have the binding affinity with their receptors regulated by post-translational modification
- IDPs adapt many different structures *in vivo* according to the cell's conditions, creating a structural or conformational ensemble
Some prediction problems

- Secondary structure of proteins
- Tertiary (3D) structure of proteins
- Docking between proteins, DNA, and/or small molecules
- Intrinsically disordered proteins/segments
- Solvent accessibility (the surface area of a protein that is accessible to a solvent, e.g., water)
Deep architectures for protein contact map prediction

Pietro Di Lena1,2, Ken Nagasta1,2 and Pierre Baldi1,2,*

1Department of Computer Science, University of California, Irvine, CA 92697, USA and 2Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, USA

ABSTRACT

Motivation: Residue-residue contact prediction is important for protein structure prediction and other applications. However, the accuracy of current contact prediction methods barely exceeds 20% on long-range contacts, falling short of the level required for ab initio structure prediction.

Results: Here, we develop a novel machine learning approach for protein contact map prediction using three steps of increasing resolution. First, we use 2D recursive neural networks to predict coarse contacts and secondary structure elements. Second, we use an energy-based method to align secondary structure elements and predict contact probabilities between residues in contacting alpha-helices or strands. Third, we use a deep neural network architecture to organize and progressively refine the prediction of contacts, integrating information from both space and time. We train the architecture on a large set of non-redundant and non-homologous domains, as well as on the test set of protein domains used for contact prediction in the two most recent CASP9 and CASP10 experiments. For long-range contacts, the accuracy of the new CMAPro predictor is close to 30%, a significant increase over existing approaches.

BIOINFORMATICS

Structural bioinformatics

Deep architectures for protein contact map prediction

Pietro Di Lena1,2, Ken Nagasta1,2 and Pierre Baldi1,2,*

1Department of Computer Science, University of California, Irvine, CA 92697, USA and 2Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, USA

ABSTRACT

Motivation: Residue-residue contact prediction is important for protein structure prediction and other applications. However, the accuracy of current contact prediction methods barely exceeds 20% on long-range contacts, falling short of the level required for ab initio structure prediction.

Results: Here, we develop a novel machine learning approach for protein contact map prediction using three steps of increasing resolution. First, we use 2D recursive neural networks to predict coarse contacts and secondary structure elements. Second, we use an energy-based method to align secondary structure elements and predict contact probabilities between residues in contacting alpha-helices or strands. Third, we use a deep neural network architecture to organize and progressively refine the prediction of contacts, integrating information from both space and time. We train the architecture on a large set of non-redundant and non-homologous domains, as well as on the test set of protein domains used for contact prediction in the two most recent CASP9 and CASP10 experiments. For long-range contacts, the accuracy of the new CMAPro predictor is close to 30%, a significant increase over existing approaches.

BIOINFORMATICS

Structural bioinformatics

Deep architectures for protein contact map prediction

Pietro Di Lena1,2, Ken Nagasta1,2 and Pierre Baldi1,2,*

1Department of Computer Science, University of California, Irvine, CA 92697, USA and 2Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, USA

ABSTRACT

Motivation: Residue-residue contact prediction is important for protein structure prediction and other applications. However, the accuracy of current contact prediction methods barely exceeds 20% on long-range contacts, falling short of the level required for ab initio structure prediction.

Results: Here, we develop a novel machine learning approach for protein contact map prediction using three steps of increasing resolution. First, we use 2D recursive neural networks to predict coarse contacts and secondary structure elements. Second, we use an energy-based method to align secondary structure elements and predict contact probabilities between residues in contacting alpha-helices or strands. Third, we use a deep neural network architecture to organize and progressively refine the prediction of contacts, integrating information from both space and time. We train the architecture on a large set of non-redundant and non-homologous domains, as well as on the test set of protein domains used for contact prediction in the two most recent CASP9 and CASP10 experiments. For long-range contacts, the accuracy of the new CMAPro predictor is close to 30%, a significant increase over existing approaches.
DeepQA: improving the estimation of single protein model quality with deep belief networks
Renzi Cao, Debwina Bhattacharya, Jie Hou and Jianlin Cheng

Abstract
Background: Protein quality assessment (QA) useful for ranking and selecting protein models has long been viewed as one of the major challenges for protein tertiary structure prediction. Especially, estimating the quality of a single protein model, which is important for selecting a few good models out of a large model pool consisting of mostly low-quality models, is still a largely unsolved problem.

Results: We introduce a novel single-model quality assessment method DeepQA based on deep belief network that utilizes a number of selected features describing the quality of a model from different perspectives, such as energy, physicochemical characteristics, and structural information. The deep belief network is trained on several large datasets consisting of models from the Critical Assessment of Protein Structure Prediction (CASP) experiments, several publicly available datasets, and models generated by our in-house ab initio method. Our experiments demonstrate that deep belief network has better performance compared to Support Vector Machines and Neural Networks on the protein model quality assessment problem, and our method DeepQA achieves the state-of-the-art performance on CASP11 dataset. It also outperformed two well-established methods in selecting good outlier models from a large set of models of mostly low quality generated by ab initio modeling methods.

Conclusion: DeepQA is a useful deep learning tool for protein single model quality assessment and protein structure prediction. The source code, executable, document and training/test datasets of DeepQA for Linux is freely available to non-commercial users at http://cactus.net.missouri.edu/DeepQA/.

Structural bioinformatics

Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks
Jack Hanson, Yuedong Yang, Kuldip Paliwal and Yaoqi Zhou

Abstract
Motivation: Capturing long-range interactions between structural but not sequence neighbors of proteins is a long-standing challenging problem in bioinformatics. Recently, long short-term memory (LSTM) networks have significantly improved the accuracy of speech and image classification problems by remembering useful past information in long sequential events. Here, we have implemented deep bidirectional LSTM recurrent neural networks in the problem of protein intrinsic disorder prediction.

Results: The new method, named SPOT-Disorder, has steadily improved over a similar method using a traditional, window-based neural network (SPINE-D) in all datasets tested without separate training on short and long disordered regions. Independent tests on four other datasets including the datasets from critical assessment of structure prediction (CASP) techniques and >10 000 annotated proteins from MultiDOB confirmed SPOT-Disorder as one of the best methods in disorder prediction. Moreover, initial studies indicate that the method is more accurate in predicting functional sites in disordered regions. These results highlight the usefulness combining LSTM with deep bidirectional recurrent neural networks in capturing non-local, long-range interactions for bioinformatics applications.
RaptorX-Property: a web server for protein structure property prediction

Sheng Wang1,2,*, Wei Li1,*, Shiwang Liu1 and Jinbo Xu1,2

1Toyota Technological Institute at Chicago, Chicago, IL, USA, 2Department of Human Genetics, University of Chicago, Chicago, IL, USA and 3School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Zhejiang, China

Received February 19, 2016; Revised April 11, 2016; Accepted April 12, 2016

ABSTRACT
RaptorX Property (http://raptorx2.uchicago.edu/StructurePropertyPred/predict) is a web server predicting structure property of a protein sequence without using any templates. It outperforms other servers, especially for proteins without close homologs in PDB or with very sparse sequence profile (i.e., carries little evolutionary information). This server employs a powerful in-house deep learning model DeepCNF (Deep Convolutional Neural Fields) to predict secondary structure (SS), solvent accessibility (ACC) and disorder regions (DISO). DeepCNF not only models complex sequence-structure relationship by a deep hierarchical architecture, but also interdependency between adjacent property labels. Our experimental results show that, tested on CASP10, CASP11 and the other benchmarks, this server can obtain ~84% Q3 accuracy for 3-state SS, ~72% Q3 accuracy for 8-state SS, ~68% Q3 accuracy for 3-state solvent accessibility, and ~0.89 area under the ROC curve (AUC) for disorder prediction.

Deep learning with feature embedding for compound-protein interaction prediction

Fangping Wan
wfpl5@mails.tsinghua.edu.cn

Jianyang (Michael) Zeng*
zengjy@gmail.com

Institute for Interdisciplinary Information Sciences, Tsinghua University

Abstract

Accurately identifying compound-protein interactions in silico can deepen our understanding of the mechanisms of drug action and significantly facilitate the drug discovery and development process. Traditional similarity-based computational models for compound-protein interaction prediction rarely exploit the latent features from current available large-scale unlabelled compound and protein data, and often limit their usage on relatively small-scale datasets. We propose a new scheme that combines feature embedding (a technique of representation learning) with deep learning for predicting compound-protein interactions. Our method automatically learns the low-dimensional implicit but expressive features for compounds and proteins from the massive amount of unlabelled data. Combining effective feature embedding with powerful deep learning techniques, our method provides a general computational pipeline for accurate compound-protein interaction prediction, even when the interaction knowledge of compounds and proteins is entirely unknown. Evaluations on current large-scale databases of the measured compound-protein affinities, such as ChEMBL and BindingDB, as well as known drug-target interactions from DrugBank have demonstrated the superior prediction performance of our method, and suggested that it can offer a useful tool for drug development and drug repositioning.
Transcription
  DNA Binding
  RNA Binding
  Enhancers/Promoters
  Alternative Splicing
  MicroRNA post-transcriptional regulation

DNA

- DNA is a *double stranded* chain of sugar molecules and phosphate residues
- Each sugar molecule contains five carbon atoms (labeled 1’ through 5’)
- Backbone bonds are between the 3’ carbon and the 5’ carbon
- Orientation of DNA is by convention 5’ to 3’
DNA

- Attached to the 1’ we can have one of four possible bases: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T)
- A,G are *purines*
- C,T are *pyrimidines*
- *Nucleotide* = sugar + phosphate + base
- DNA can reach in the 100s of millions of base pairs
RNA

- Single stranded
- Uracil (U) instead of thymine (T)
- Different types of RNA
  - mRNA (messenger RNA)
  - tRNA (transfer RNA)
  - rRNA (ribosomal RNA)
  ... and recently discovered ncRNA in the “RNAi world”: miRNA, siRNA, snoRNA, stRNA, snRNA
- RNA is much less stable than DNA

Central Dogma
Genes

- **Gene**: a segment of DNA which encodes for at least one polypeptide chain (usually mRNA)
- It includes regions preceding and following the coding region (UTR) and intervening sequences (**introns**)
- Genes usually lie in non-repetitive DNA

Transcription

- The synthesis of mRNA on a DNA template
- **RNA polymerase** is the enzyme that catalyzes this process (**pol II** in eukaryotes)
- RNA polymerase transcribes 1Kbps/sec
- The first base pair transcribed is called **transcription start site** (TSS)
**Figure 1.29** The gene may be longer than the sequence coding for protein.

**Figure 2.10** Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.
Alternative splicing

Transcription

- **Promoter**: a region of DNA involved in binding of RNA polymerase to initiate transcription
- **Enhancer**: a region of DNA that increases the utilization of (some) promoters (it can function in either orientations and any location relative to the promoter)
- **Repressor**: a region of DNA that decreases the utilization of (some) promoters
Promoters and Enhancers

Transcription control

- Different factors are involved in the transcription machinery
  - binding of transcription factors to DNA
  - ability of DNA to bend
  - relative location of the binding sites
  - interaction between transcription factors
  - DNA methylation, nucleosomes (epigenetics)
  - presence \( CpG \) islands ("p" is for phosphate)
  - ...
Genetic “circuits”

- If C then D
- If B then NOT D
- If A and B then D
- If D then B
- Gene D
- Gene B

Slide by Serafim Batzoglou, Stanford U.
Example: A Human heat shock protein

- TATA box: positioning transcription start
- TATA, CCAAT: constitutive transcription
- GRE: glucocorticoid response el.
- MRE: metal response element
- HSE: heat shock element

Slide by Serafim Batzoglou, Stanford U.

Translation

- The synthesis of a protein on the mRNA template
- Takes place inside ribosomes
- Ribosomes are made of rRNA
- Ribosomes translate about 60 bases/sec (<0.0001% error rate)
- mRNA is translated into the corresponding amino acids by ribosomes + tRNA
**Figure 1.30** Gene expression is a multistage process.

**Figure 5.8** A polyribosome consists of an mRNA being translated simultaneously by several ribosomes moving in the direction from 5' to 3'. Each ribosome has two tRNA molecules: one carrying the nascent protein, the second carrying the next amino acid to be added.
Non-coding RNAs

- **dsRNA**: double stranded RNA, typically longer than 30 nt
- **miRNA**: microRNA, 21-25 bases
  - Encoded by endogenous (‘within’) genes
  - Hairpin precursors
  - Recognize multiple targets
- **siRNA**: short-interfering RNA, 21-25 bases
  - Mostly exogenous origin
  - dsRNA precursors
  - May be target specific
Some prediction problems

- DNA/protein binding (e.g., TFs)
- (nc)RNA/protein binding
- miRNAs
- Alternative splicing (isoforms)
- Gene expression or circadian rhythms of gene expression
- Enhancers and promoters location
- Transcript boundaries (start gene, end gene)
- Translation initiation site
DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins

Hamid Reza Hassananzadeh
Department of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
Email: hassananzadeh@gatech.edu

May D. Wang
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, Georgia 30332
Email: maywang@bme.gatech.edu

Abstract—Transcription factors (TFs) are macromolecules that bind to complimentary specific subregions of DNA promoters and initiate transcription. Finding the exact location of these binding sites (aka motifs) is important in a variety of domains such as drug design and development. To address this need, several in vivo and in vitro techniques have been developed so far that try to characterize and predict the binding specificity of a protein to different DNA loci. The major problem with these techniques is that they are not accurate enough in prediction of both the binding affinity and characterization of the corresponding motifs. As a result, downstream analysis is required to uncover the locations where proteins of interest bind. Here, we propose DeeperBind, a long short-term recurrent convolutional network for prediction of protein binding specificities with respect to DNA probes. DeeperBind can model the positional dynamics of probe sequences and hence recons with the contributions made by individual subregions in DNA sequences, in an effective way. Moreover, it can be trained and tested on datasets containing varying-length sequences. We apply our pipeline to the datasets derived from protein binding microarrays (PBMs), an in-vita high-throughput technology for quantification of protein-DNA binding preferences, and present promising results. To the best of our knowledge, this is the most accurate pipeline that can predict binding specificities of DNA sequences from the data produced by high-throughput technologies through utilization of the power of deep learning for feature generation and positional dynamics modeling.

Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks

Shashank Singh1, Yang Yang1, Barnabás Póczos1, and Jian Ma1,2,∗

1Machine Learning Department
2Computational Biology Department
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
∗Corresponding author: jianma@cs.cmu.edu

Abstract

In the human genome, distal enhancers are involved in regulating target genes through proximal promoters by forming enhancer-promoter interactions. However, although recently developed high throughput experimental approaches have allowed us to recognize potential enhancer-promoter interactions genome-wide, it is still largely unknown whether there are sequence-level instructions encoded in our genome that help govern such interactions. Here we report a new computational method (named “SPEID”) using deep learning models to predict enhancer-promoter interactions based on sequence-based features only, when the locations of putative enhancers and promoters in a particular cell type are given. Our results across six different cell types demonstrate that SPEID is effective in predicting enhancer-promoter interactions as compared to state-of-the-art methods that use non-sequence features from functional genomic signals. This work shows for the first time that sequence-based features alone can reliably predict enhancer-promoter interactions genome-wide, which provides important insights into the sequence determinants for long-range gene regulation.
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi1,2, Andrew Delong1,6, Matthew T Weirau1-3 & Brendan J Frey1-3

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 DeepBind and have built a stand-alone software tool that is fully automatic and handles millions of sequences per experiment. Specificities determined by DeepBind are readily visualized as a weighted ensemble of position weight matrices as a ‘mutation map’ that indicates how variations affect binding within a specific sequence.

A deep learning framework for modeling structural features of RNA-binding protein targets

Sai Zhang1, Jinglin Zhou2, Hailin Hu1, Halping Gong1, Ligong Chen1, Chao Cheng1 and Jianying Zeng1

1 Institute for Humaniora Information Sciences, Tsinghua University, Beijing 100084, China; 2 Department of Pharmacology and Pharmaceutical Sciences, School of Medicine, Tsinghua University, Beijng 100084, China; 3 School of Life Sciences, Tsinghua University, Beijing 100084, China; 4 Biomedical Key Laboratory of Biomechanics, Tsinghua University, Beijing 100084, China and 5 Department of Genetics, Institute for Quantitative Biomedical Sciences, Nova Center for Cancer Center, School of Medicine at Dartmouth, Hanover, NH 03755, USA

ABSTRACT

RNA-binding proteins (RBPs) play important roles in the post-transcriptional control of RNA. Identifying RBP binding sites and characterizing RBP binding preferences are key steps toward understanding the basic mechanisms of the post-transcriptional regulation. Over the past decade, computational methods have been developed for modeling RBP binding preferences, discussing a complete structural representation of the RBP targets by integrating their available functional features in all three dimensions is still a challenging task. In this paper, we develop a general and flexible deep learning framework for modeling structural binding preferences and predicting binding sites of RBPs, which takes (predicted) RNA tertiary structural information into account for the first time. Our framework constructs a unified representation that characterizes the structural specificities of RBP targets in all three dimensions. To verify our model, we tested our method on several classes of RNA-binding proteins and distal novel candidate binding sites and discover potential binding motifs. Through testing on the real RBP structures, we have demonstrated that our deep learning framework can automatically extract effective hidden structural features from the encoded raw sequence and structural profiles, and predict accurate RBP binding sites. In addition, we have performed the first study to show that integrating the additional RNA tertiary structural features can improve the model performance in predicting RBP binding sites, especially for the pppGpp/ADP ribosyltransferase binding protein (Ptp), which also provides a new evidence to support the view that RBPs may have specific tertiary structural binding preferences. In summary, the results of this study demonstrate that using the RNA tertiary structures and distal binding sites yield satisfactory results with experimental support from the literature and further demonstrate the feasibility of incorporating RNA tertiary structural information into the prediction model. The source code of our approach can be found at https://github.com/huqinjin/deepLBP.

INTRODUCTION

RNA-binding proteins (RBPs) play important roles in various cellular processes, such as alternative splicing, RNA editing, and mRNA localization and translation (1). RBPs contain several core RNA-binding domains (RBDs), e.g., the RNA recognition motif (RRM), the leucine zipper-containing domain (LZD), and the zinc-finger motif (ZNF). The RBDs are responsible for binding to specific RNA sequences, and each RBD has a unique structure and structural properties (2). Although it has been shown that several important disease-related diseases, cancer and cardiovascular diseases, can be caused by the dysfunctions of certain RBPs (e.g., misregulated splicing or translation), the mechanisms of these diseases are still not fully understood (3).

Recently, the advent of high-throughput experimental methods, such as the cross-linking immunoprecipitation (CLIP) assay, has greatly advanced the generation wide studies of RNA-protein interactions (4). Despite the progress of these experimental techniques, the collected data still suffer from the false-positive and false-negative problems due...
A Deep Boosting Based Approach for Capturing the Sequence Binding Preferences of RNA-Binding Proteins from High-Throughput CLIP-Seq Data

Shuoa Li1, Fangzeng Dong2, Yunyan Wu3,4, Hai Zhang3, Chen Zhang3, Xiao Liu4, Tao Jiang4, and Ruiyang Zeng2

1School of Life Sciences, Tsinghua University, Beijing, China.
2Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
3Present address: Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA.
4Department of Computer Science and Engineering, University of California, Riverside, CA.
5MOE Key Lab of Bioinformatics and Biometric Information, TMU/T.
6Department of Computer Science and Technology, Tsinghua University, Beijing, China.
7To whom correspondence should be addressed. Email: mengji21@tsinghua.edu.cn.
8These authors contributed equally to this work.

Abstract
Characterizing the binding behavior of RNA-binding proteins (RBPs) is important for understanding their functional roles in gene expression regulation. However, current high-throughput experimental methods for identifying RBPs, such as CLIP-seq and RNAcompete, usually suffer from the false positive and false negative issues. Here, we develop a deep boosting based machine learning approach, called DeepBooster, to accurately model the binding sequence preferences and identify the corresponding binding targets of RBPs from CLIP-seq data. Comprehensive validation tests have shown that DeepBooster can outperform other state-of-the-art approaches in predicting RBP targets and recover false negatives that are common in current CLIP-seq data. In addition, we have demonstrated several new potential applications of DeepBooster in understanding the regulatory frame...

Integrative Deep Models for Alternative Splicing
Anupama Jha1, Matthew R. Gazzara1,2,3 and Yoseph Barash1,2,*

January 31, 2017

1Department of Computer and Information Science, School of Engineering, and
2Department of Genomics, and
3Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.
*Correspondence should be addressed to yosephb@upenn.edu

Abstract
Advancements in sequencing technologies have highlighted the role of alternative splicing (AS) in increasing transcriptome complexity. This role of AS, combined with the relation of aberrant splicing to malignant states, motivated two streams of research, experimental and computational. The first involves a myriad of techniques such as RNA-Seq and CLIP-Seq to identify splicing regulators and their putative targets. The second involves probabilistic models, also known as splicing codes, which infer regulatory mechanisms and predict splicing outcome directly from genomic sequence. To date, these models have utilized only expression data. In this work we address two related challenges: Can we improve on previous models for AS outcome prediction and can we integrate additional sources of data to improve predictions for AS regulatory factors. We perform a detailed comparison of two previous modeling approaches, Bayesian and Deep Neural networks, dissecting the confounding effects of datasets and target functions. We then develop a new target function for AS prediction and show that it significantly improves model accuracy. Next, we develop a modeling framework to incorporate CLIP-Seq, knockdown and over-expression experiments, which are inherently noisy and suffer from missing values. Using several datasets involving key splicing factors in mouse brain, muscle and heart we demonstrate both the prediction improvements and biological insights offered by our new models. Overall, the framework we propose offers a scalable integrative solution to improve splicing code modeling as vast amounts of relevant genomic data become available. Availability: code and data will be available on Github following publication.
Deep learning of the tissue-regulated splicing code

Michael K. K. Leung$^{1,2}$, Hui Yuan Xiong$^{1,2}$, Leo J. Lee$^{1,2}$ and Brendan J. Frey$^{1,2,3,*}$

$^1$Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario M5S 3G4, $^2$Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada and $^3$Canadian Institute for Advanced Research, Toronto, Ontario M5G 1Z8, Canada

ABSTRACT

Motivation: Alternative splicing (AS) is a regulated process that directs the generation of different transcripts from single genes. A computational model that can accurately predict splicing patterns based on genomic features and cellular context is highly desirable, both in understanding this widespread phenomenon, and in exploring the effects of genetic variations on AS.

Methods: Using a deep neural network, we developed a model inferred from mouse RNA-Seq data that can predict splicing patterns in individual tissues and differences in splicing patterns across tissues. Our architecture uses hidden variables that jointly represent features in genomic sequences and tissue types when making predictions. A graphics processing unit was used to greatly reduce the training time of our models with millions of parameters.

Results: We show that the deep architecture surpasses the performance of the previous Bayesian method for predicting AS patterns. With the proper optimization procedure and selection of hyperparameters, we demonstrate that deep architectures can be beneficial, even with a moderately sparse dataset. An analysis of what the model has learned in terms of the genomic features is presented.

Previously, a ‘splicing code’ that uses a Bayesian neural network (BNN) was developed to infer a model that can predict the outcome of AS from sequence information in different cellular contexts (Xiong et al., 2011). One advantage of Bayesian methods is that they protect against overfitting by integrating over models. When the training data are sparse, as is the case for many datasets in the life sciences, the Bayesian approach can be beneficial. It was shown that the BNN outperforms several common machine learning algorithms, such as multinomial logistic regression (MLR) and support vector machines, for AS prediction in mouse trained using microarray data.

There are several practical considerations when using BNNs. They often rely on methods like Markov Chain Monte Carlo (MCMC) to sample models from a posterior distribution, which can be difficult to speed up and scale up to a large number of hidden variables and a large volume of training data. Furthermore, computation-wise, it is relatively expensive to get predictions from a BNN, which requires computing the average predictions of many models.

Recently, deep learning methods have surpassed the state-of-

Integrative Deep Models for Alternative Splicing

Anupama Jha$^1$, Matthew R. Gazzara$^{1,2,3}$ and Yoseph Barash$^{1,2,*}$

January 31, 2017

$^1$Department of Computer and Information Science, School of Engineering, and $^2$Department of Genetics, and $^3$Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

*Correspondence should be addressed to yosephb@upenn.edu

Abstract

Advancements in sequencing technologies have highlighted the role of alternative splicing (AS) in increasing transcriptome complexity. This role of AS, combined with the relation of aberrant splicing to malignant states, motivated two streams of research, experimental and computational. The first involves a myriad of techniques such as RNA-Seq and CLIP-Seq to identify splicing regulators and their putative targets. The second involves probabilistic models, also known as splicing codes, which infer regulatory mechanisms and predict splicing outcome directly from genomic sequence. To date, these models have utilized only expression data. In this work we address two related challenges: Can we improve on previous models for AS outcome prediction and can we integrate additional sources of data to improve predictions for AS regulatory factors? We perform a detailed comparison of two previous modeling approaches, Bayesian and Deep Neural networks, dissecting the confounding effects of datasets and target functions. We then develop a new target function for AS prediction and show that it significantly improves model accuracy. Next, we develop a modeling framework to incorporate CLIP-Seq, knockdown and over-expression experiments, which are inherently noisy and suffer from missing values. Using several datasets involving key splice factors in mouse brain, muscle and heart we demonstrate both the prediction improvements and biological insights offered by our new models. Overall, the framework we propose offers a scalable integrative solution to improve splicing code modeling as vast amounts of relevant genomic data become available.

Availability: Code and data will be available on Github following publication.
Gene expression

Gene expression inference with deep learning

Yifei Chen1,4, 1, Yi Li1, 1, Rajiv Narayan1, Aravind Subramanian2 and Xiaohui Xie1, 3, 4

1Department of Computer Science, University of California, Irvine, CA 92697, USA, 2Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, 3Center for Complex Biological Systems, University of California, Irvine, CA 92697, USA and 4Beijing Research-Big Data Lab, Beijing, 100193, China

*To whom correspondence should be addressed.

The authors wish to be known that, in their opinion, the first two authors should be regarded as joint first Authors.

Accepted 8 September 2016; revised on December 14, 2016; accepted on February 2, 2017.

Abstract

Motivation: Large-scale gene expression profiling has been widely used to characterize cellular states in response to various disease conditions, genetic perturbations, etc. Although the cost of whole-genome expression profiles has been dropping steadily, generating a comprehensive expression profiling over thousands of samples is still very expensive. Recognizing that gene expressions are often highly correlated, researchers from the NH-LINCE program have developed a cost-effective strategy of profiling only 1000 carefully selected landmark genes and relying on computational methods to infer the expression of remaining target genes. However, the computational approach adopted by the LINCE program is currently based on linear regression (LR), limiting its accuracy since it does not capture complex nonlinear relationships between expressions of genes.

Results: We present a deep learning method (abbreviated as D-GEK) to infer the expression of target genes from the expression of landmark genes. We used the microarray-based Gene Expression Omnibus dataset, consisting of 113K expression profiles, to train our model and compare its performance to those from other methods. In terms of mean absolute error averaged across all genes, deep learning significantly outperforms LR with 15.33% relative improvement. A gene-wise comparative analysis shows that deep learning achieves lower error than LR in 98.91% of the target genes. We also tested the performance of our learned model on an independent RNA-Seq-based GTEx dataset, which consists of 9381 expression profiles. Deep learning still outperforms LR with 6.57% relative improvement, and achieves lower error in 81.31% of the target genes.

What time is it? Deep learning approaches for circadian rhythms

Forest Agostinelli1, 3, Nicholas Ceglia1, Babak Shahbaba2, Paolo Sassone-Corsi2 and Pierre Baldi1, 3

1Department of Computer Science, 2Department of Statistics and 3Department of Biological Chemistry, University of California-Irvine, Irvine, CA 92697, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Circadian rhythms date back to the origins of life, are found in virtually every species and every cell, and play fundamental roles in functions ranging from metabolism to cognition. Modern high-throughput technologies allow the measurement of concentrations of transcripts, metabolites and other species along the circadian cycle creating novel computational challenges and opportunities, including the problems of inferring whether a given species oscillate in circadian phases or not, and inferring the time at which a set of measurements was taken.

Results: We first curate several large synthetic and biological time series datasets containing labels for both periodic and aperiodic signals. We then use deep learning methods to develop and train BIO_CYCLE, a system to robustly estimate which signals are periodic in high-throughput circadian experiments, producing estimates of amplitudes, periods, phases, as well as several statistical significance measures. Using the curated data, BIO_CYCLE is compared to other approaches and shown to achieve state-of-the-art performance across multiple metrics. We then use deep learning methods to develop and train BIO_CLOCK to robustly estimate the time at which a particular single-time-point transcriptomic experiment was carried. In most cases, BIO_CLOCK can reliably predict time, within approximately 1% of the expression levels of only a small number of core clock genes. BIO_CLOCK is shown to work reasonably well across tissue types, and often with only small degradation across conditions. BIO_CLOCK is used to annotate most mouse experiments found in the GEO database with an inferred time stamp.
Genome-Wide Prediction of cis-Regulatory Regions Using Supervised Deep Learning Methods

Yifeng Li1,2, Wenguang Shi1, and Wyeth W. Wasserman∗1

1Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada
2Information and Communication Technologies, National Research Council Canada, Ottawa, Ontario, Canada

Abstract

Identifying active cis-regulatory regions in the human genome is critical for understanding gene regulation and assessing the impact of genetic variation on phenotype. Based on rich data resources such as the Encyclopedia of DNA Elements (ENCODE) and the Functional Annotation of the Mammalian Genome (FANTOM) projects, we introduce DECRG, the first supervised deep learning approach for the identification of enhancer and promoter regions in the human genome. Due to their ability to discover patterns in large and complex data, the introduction of deep learning methods enables a significant advance in our knowledge of the genomic locations of cis-regulatory regions. Using models for well-characterized cell lines, we identify key experimental features that contribute to the predictive performance. Applying DECRG, we delineate locations of 300,000 candidate enhancers genome wide (6.8% of the genome), of which 40,000 are supported by bidirectional transcription data and 26,000 candidate promoters (0.6% of the genome).

IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction

Xiaoyong Fan1,2, Yang Xian Fan3,4, Junchi Yan3 and Hong-Bin Shen5

Abstract

Background: Non-coding RNAs (ncRNAs) play crucial roles in many biological processes, such as post-transcriptional gene regulation. ncRNAs mainly function through interaction with RNA-binding proteins (RBPs). To understand the function of ncRNAs, a fundamental step is to identify which protein is involved in its interaction. Therefore it is promising to computationally predict RBPs, where the major challenge is that the interaction pattern or motif is difficult to be found.

Results: In this study, we propose a computational method: IPMiner (Interaction Pattern Miner) to predict ncRNA-protein interactions from sequences, which makes use of deep learning and further improves its performance using stacked ensembling. One of the IPMiner's typical merits is that it is able to mine the hidden sequential interaction patterns from sequence composition features of protein and RNA sequences using stacked autoencoder, and then the learned hidden features are fed into random forest models. Finally, stacked ensembling is used to integrate different predictors to further improve the prediction performance. The experimental results indicate that IPMiner achieves superior performance on the tested ncRNA-protein interaction dataset with an accuracy of 0.89, AUC of 0.85, specificity of 0.81, precision of 0.83, and Matthews correlation coefficient of 0.84, respectively. We further comprehensively investigate IPMiner on other RNA-protein interaction datasets, which yields better performance than the state-of-the-art methods, and the performance has an increase of over 30% on some tested benchmark datasets. In addition, we further apply IPMiner for large-scale prediction of ncRNA-protein network, that achieves promising prediction performance.

Conclusions: By integrating deep neural network and stacked ensembling, from simple sequence composition features, IPMiner can automatically learn high-level abstraction features, which has strong discriminant ability for RNA-protein detection. IPMiner achieved high performance on our constructed ncRNA-protein benchmark dataset and other RNA-protein datasets. IPMiner tool is available at http://www.cs.ubc.ca/labs/TML/IPMiner.
Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters

Yifeng Li, Chih-Yu Chen, and Wyeth W. Waterman

Centre for Molecular Medicine and Therapeutics, University of British Columbia, 501 West 20th Avenue, Vancouver, BC V5Z 4H4, Canada
lyifeng, chihyuc, wyeth@cmmt.ubc.ca

Abstract. Sparse linear models approximate target variable(s) by a sparse linear combination of input variables. The sparseness is realized through a regularization term. Since they are simple, fast, and able to select features, they are widely used in classification and regression. Essentially linear models are shallow feed-forward neural networks which have three limitations: (1) incompatibility to model non-linearity of features, (2) inability to learn high-level features, and (3) unnatural extensions to select features in multi-class case. Deep neural networks are models structured by multiple hidden layers with non-linear activation functions. Compared with linear models, they have two distinctive strengths: the capability to (1) model complex systems with non-linear structures, (2) learn high-level representation of features. Deep learning has been applied in many large and complex systems where deep models significantly outperform shallow ones. However, feature selection at the input level, which is very helpful to understand the nature of a complex system, is still not well-studied. In genome research, the cis-regulatory elements in non-coding DNA sequences play a key role in the expression of genes. Since the activity of regulatory elements involves highly interactive factors, a deep tool is strongly needed to discover informative features. In order to address the above limitations of shallow and deep models for selecting features of a complex system, we propose a deep feature selection model that (1) takes advantages of deep models to model non-linearity and (2) conveniently selects a subset of features right at the input level for multi-class data. We applied this model to the identification of active enhancers and promoters by integrating multiple sources of genomic information. Results show that our model outperforms elastic net in terms of size of discriminative feature subset and classification accuracy.

Published online 1 November 2014
doi:10.1038/nrg3505

DEEP: a general computational framework for predicting enhancers

Dimitrios Kifelogiannis1, Panos Kalitsis1 and Vladimir B. Bajic2

1Computation, Electrical and Mathematical Sciences and Engineering Division (CEMES), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia and Computational Biology Program, Molecular and Medical Anthropology Program, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
2Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843-3112, USA

ABSTRACT

Transcription regulation in multicellular eukaryotes is orchestrated by a number of DNA functional elements located at gene regulatory regions. Some regulatory regions (e.g. enhancers) are located far away from the gene they affect, identification of distal regulatory elements is a challenge for the biotechnomic research. Although existing methodologies increased the number of computationally predicted enhancers, performance inconsistency of computational models across different cell lines, cell imbalance within the learning sets and ad hoc rules for selecting enhancer candidates for supervised learning, are some key questions that require further exploration. In this study we developed DEEP, a deep learning framework for model computational enhancer prediction that integrates computational models with divergent characteristics that stimulate the analysis of enhancer properties in a wide variety of cellular conditions. In our method we train many shallow computational models that we combine to classify DNA regions as enhancers or non-enhancers. DEEP uses features obtained from histone modification marks or attributes coming from sequence characteristics. Experimental results indicate that DEEP performs better than four state-of-the-art methods in the Broad dataset. We report the first computational enhancer prediction results on Fantom5 data, where DEEP achieves 96.7% accuracy and 99.4% geometric mean (G-mean) of specificity and sensitivity across 26 different tissues. We also report deep learning framework for enhancer data from VISTA database, DEEP-VISTA, when tested on an independent test set, achieved 80.1% of accuracy and 88.5% of precision. DEEP framework is publicly available at http://vista.usc.edu/deep/.

INTRODUCTION

Transcription regulation in human genes is a complex process. (1) Promoters and cis-regulatory elements, which are organized into a combinatorial network, functionally determine gene expression levels and tissue specificity. However, the boundaries of these regulatory elements are not always spatially proximal to the target gene. This implies that they may not always be co-expressed, or have different functions, depending on the cellular state (e.g. are we active in defense, or are we stress-enhanced or not?). This is not yet fully understood. (2) Recent advances in high-throughput experiments like the ENCODE project indicate that interactions between proximal and distal regulatory elements orchestrate gene expression between different cell lines. The distance between enhancers and promoters is not necessarily correlated with their expression profile. Domain assembling and the role of enhancers in the cascade from enhancer to transcription is not yet fully understood. (3) ENCODE data reveal that enhancers do not always overlap with the accessible regions of the gene (i.e. regulatory elements are not spatially proximal to the gene). Enhancer accessibility is a complex process that is not yet fully understood. (4) In many cases, enhancers can have different function depending on the cell line (e.g. are we active in defense, or are we stress-enhanced or not?), and their functional mechanisms is not yet fully known. In this line with (1) we characterize enhancers as co-acting DNA regulatory elements in the regulation of the target gene. Enhancers activate gene transcription by recruiting transcription factors (TFs) and their complexes. For the models, enhancer regions frequently contain dozens of binding sites of various TFs. This makes integrating different regions and their effects on the enhancer output very challenging.
DeepTarget: End-to-end Learning Framework for microRNA Target Prediction using Deep Recurrent Neural Networks

Byunghan Lee
Electrical and Computer Eng.
Seoul National University
Seoul 08826, Korea

Seunghyun Park
Electrical and Computer Eng.
Seoul National University
Seoul 08826, Korea

Junghwan Baek
Interdisciplinary Program in Bioinformatics
Seoul National University
Seoul 08826, Korea

Sungroh Yoon
Electrical and Computer Eng.
& Interdisciplinary Program in Bioinformatics
Seoul National University
Seoul 08826, Korea
sryoon@snu.ac.kr

ABSTRACT
MicroRNAs (miRNAs) are short sequences of ribonucleic acids that control the expression of target messenger RNAs (mRNAs) by binding them. Robust prediction of miRNA-mRNA pairs is of utmost importance in deciphering gene regulation but has been challenging because of high false positive rates, despite a deluge of computational tools that normally require laborious manual feature extraction. This paper presents an end-to-end machine learning framework for miRNA target prediction. Leveraged by deep recurrent neural networks based auto-encoding and sequence-sequential interaction learning, our approach not only delivers an unprecedented level of accuracy but also eliminates the need for manual feature extraction. The performance gap between the proposed method and existing alternatives is substantial (over 25% increase in F-measure), and our approach delivers a quantum leap in the longstanding challenge of robust miRNA target prediction. [availability: http://data.snu.ac.kr/deepTarget]

Keywords
microRNA, deep learning, recurrent neural networks, LSTM

1. INTRODUCTION
MicroRNAs (miRNAs) are small non-coding RNA molecules that can control the function of their target messenger RNAs (mRNAs) by down-regulating the expression of the targets [4]. By controlling the gene expression at the RNA level, miRNAs are known to be involved in various biological processes and diseases [27]. As miRNAs play a central role in the post-transcriptional regulation of more than 60% of protein coding genes [11], investigating miRNAs is of utmost importance in many disciplines of life science. As explained further in Section 2.3.1, miRNAs are derived from the precursor miRNAs (pre-miRNAs) and then exhibit their regulatory function by binding to the target sites present in miRNAs. Two types of computational problems about miRNAs have naturally arisen in bioinformatics: miRNA seed identification (i.e., the problem of locating the genes that encode

Deep modeling of gene expression regulation in an Erythropoiesis model

Olgert Denas
ODENAS@EMORY.EDU
Department of Mathematics and Computer Science, 400 Dowman Dr., Atlanta, GA 30322, USA

James Taylor
JAMES.TAYLOR@EMORY.EDU
Department of Mathematics and Computer Science, 400 Dowman Dr., Atlanta, GA 30322, USA
Department of Biology, 1510 Clifton Road NE, Atlanta, GA 30322

Abstract
The fate of differentiation of G1E cells is determined, among other things, by a handful of transcription factors (TFs) binding the neighborhood of appropriate gene targets. The problem of understanding the dynamics of gene expression regulation is a feature learning problem on high-dimensional space determined by the size of the gene neighborhoods, but can be projected on a much lower dimensional manifold whose space depends on the number of TFs and the number of ways they interact. To learn this manifold, we train a deep convolutional network on the activity of TF binding on 2033 gene neighborhoods labeled by bloodline levels of target gene expression. After unsupervised training of the model we achieve 77% accuracy as estimated by 10-fold CV.

We discuss methods for the representation of the model, knowledge taken into the input space. We use this representation to highlight important patterns and genomic locations with biological importance.

and mouse ENCODE projects (ENCODE Project Consortium, 2012) for the generation of hundreds of assays targeting specific transcription factors (TFs) on a variety of cell lines. ChIP-Seq derived TFs correlate well with the locations of functional genome elements, however, the resolution is low and the data lacks statistical power being limited to a single cell-TF pair. It is a challenge today to effectively use the data from this technology for accurate prediction of functional elements. Data noise is bound to the technology, and more context can be used to improve accuracy if prediction models combined data from several experiments.

Here, we propose a deep convolutional architecture as candidate.

Motivated initially by the visual cortex (Hubel & Wiesel, 1962, 1965), deep convolutional architectures (LeCun et al., 1989) have been very successful predictive systems in digit classification, and image and object recognition (Bengio & LeCun, 2007; LeCun et al., 2014) and natural language processing (Collobert & Weston, 2008). A convolutional neural network (CNN) replicates feature detectors across all connections between two layers. Thus, sharing the weights
Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model

Lujia Chen, Chunhui Cai, Vicky Chen and Xinghua Lu

From The Fourteenth Asia Pacific Bioinformatics Conference (APBIC 2016)
San Francisco, CA, USA, 11 - 13 January 2016

Abstract

Background: A living cell has a complex, hierarchically-organized signaling system that encodes and assimilates diverse environmental and intracellular signals, and further transmits signals that control cellular responses, including a tightly controlled transcriptional program. An important and yet challenging task in systems biology is to reconstruct cellular signaling system in a data-driven manner. In this study, we investigate the utility of deep hierarchical neural networks in learning and representing the hierarchical organization of yeast transcriptomic machinery.

Results: We have designed a sparse autoencoder model consisting of a layer of observed variables and four layers of hidden variables. We applied the model to over a thousand of yeast microarrays to learn the encoding system of yeast transcriptomic machinery. After model selection, we evaluated whether the trained model captured biologically sensible information. We show that the latent variables in the first hidden layer correctly captured the signals of yeast transcription factors (TFs), obtaining a close-to-one-to-one mapping between latent variables and TFs. We further show that genes regulated by latent variables at higher hidden layers are often involved in a common biological process, and the hierarchical relationships between latent variables conform to existing knowledge. Finally, we show that information captured by the latent variables provide more abstract and concise representations of each microarray, enabling the identification of better separated clusters in comparison to gene-based representation.

Conclusions: Contemporary deep hierarchical latent variable models, such as the autoencoder, can be used to partially recover the organization of transcriptomic machinery.

Reverse-complement parameter sharing improves deep learning models for genomics

Avanti Shrikumar 1*, Peyton Greenside 2* and Anshul Kundaje 1,3

1Computer Science, Stanford University, Stanford, 94305, USA
2Biomedical Informatics, Stanford University, Stanford, 94305, USA
3Genetics, Stanford University, Stanford, 94305, USA
*Co-first authors

Abstract

Deep learning approaches that have produced breakthrough predictive models in computer vision, speech recognition and machine translation are now being successfully applied to problems in regulatory genomics. However, deep learning architectures used thus far in genomics are often directly ported from computer vision and natural language processing applications with few, if any, domain-specific modifications. In double-stranded DNA, the same pattern may appear identically on one strand and its reverse complement due to complementary base pairing. Here, we show that conventional deep learning models that do not explicitly model this property can produce substantially different predictions on forward and reverse-complement versions of the same DNA sequence. We present four new convolutional neural network layers that leverage the reverse-complement property of genomic DNA sequence by sharing parameters between forward and reverse-complement representations in the model. These layers guarantee that forward and reverse-complement sequences produce identical predictions within numerical precision. Using experiments on simulated and in vivo transcription factor binding data, we show that our proposed architectures lead to improved performance, faster learning and cleaner internal representations compared to conventional architectures trained on the same data.
TITER: predicting translation initiation sites by deep learning

Sai Zhang 1,1, Haolin Hu 2,1, Tao Jiang 3,4,5, Lei Zhang 2,6 and Jianyang Zeng 1,*

1Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China.
2School of Medicine, Tsinghua University, Beijing 100084, China.
3Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA.
4MOE Key Lab of Bioinformatics and Bioinformatics Database, TNUST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
5Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA.
6These authors contributed equally to this work.
*Correspondence should be addressed.
Associate Editor: XXXXXX
Revised on XXXXX; revised on XXXXX; accepted on XXXXX.

Abstract

Motivation: Translation initiation is a key step in the regulation of gene expression. In addition to the annotated translation initiation sites (TISs), the translation process may also start at multiple alternative TISs (including both AUG and non-AUG codons), which makes it challenging to predict TISs and study the underlying regulatory mechanisms. Meanwhile, the advent of several high-throughput sequencing techniques for profiling initiating ribosomes at single-nucleotide resolution, e.g., GT1-seq and GT1-seq, provides abundant data for systematically studying the general principles of translation initiation and the development of computational methods for TIS identification.

Methods: We have developed a deep learning based framework, named TITER, for accurately predicting TISs on a genome wide scale based on GT1-seq data. TITER extracts the sequence features of translation initiation from the surrounding sequence contexts of TISs using a hybrid neural network and further integrates the prior preference of TIS codon composition into a unified prediction framework.

Results: Extensive tests demonstrated that TITER can greatly outperform the state-of-the-art prediction methods in identifying TISs. In addition, TITER was able to identify important sequence signatures for individual types of TIS codons, including a Kozak-sequence like motif for AUG start codon. Furthermore, the TITER prediction score can be related to the strength of translation initiation in various biological scenarios, including the repressive effect of the upstream open reading frames (uORFs) on gene expression and the mutational effects influencing translation initiation efficiency.

Characterizing RNA Pseudouridylation by Convolutional Neural Networks

Xuan He 1, Sai Zhang 1, Yanqing Zhang 1, Tao Jiang 3,4, and Jianyang Zeng 1,*

1Institute for Interdisciplinary Information Science, Tsinghua University, Beijing 100084, China.
2Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA.
3MOE Key Lab of Bioinformatics and Bioinformatics Database, TNUST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
4Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA.
5To whom correspondence should be addressed.
E-mail: zengy321@tsinghua.edu.cn

March 13, 2017

Abstract

The most prevalent post-transcriptional RNA modification, pseudouridine (Ψ), also known as the fifth ribonucleoside, is widespread in rRNAs, tRNAs, miRNAs, siRNAs and snRNAs. Pseudouridines in RNAs are implicated in many aspects of post-transcriptional regulation, such as the maintenance of translation fidelity, control of RNA stability and stabilization of RNA structure. However, our understanding of the functions, mechanisms as well as precise distribution of pseudouridines (especially in miRNAs) still remains largely unclear. Though thousands of RNA pseudouridylate sites have been identified by high-throughput experimental techniques recently, the landscape of pseudouridines across the whole transcriptome has not yet been fully delineated. In this study, we present a highly effective model, called PULSE (Pseudouridine Sites) (http://sites.zenglab.com), to predict novel Ψ sites from large-scale profiling data of pseudouridines and characterize the contextual sequence features of pseudouridylation. PULSE employs a deep learning framework, called convolutional neural network (CNN), which has been successfully and widely used for sequence pattern discovery in the literature. Our extensive validation tests demonstrated that PULSE can complement conventional learning models and achieve high prediction accuracy, thus enabling us to further characterize the pseudouridine-wide landscape of pseudouridines sites. Overall, PULSE can provide a useful tool to further investigate the functional roles of pseudouridylation in post-transcriptional regulation.
DeepATAC: A deep-learning method to predict regulatory factor binding activity from ATAC-seq signals

Na ozumi Hiranuma\textsuperscript{1}, Scott Lundberg\textsuperscript{1}, Su-In Lee\textsuperscript{1,2}

\textsuperscript{1} Paul G. Allen School of Computer Science and Engineering,
\textsuperscript{2} Department of Genome Sciences, School of Medicine, University of Washington

Abstract

Determining the binding locations of regulatory factors, such as transcription factors and histone modifications, is essential to both basic biology research and many clinical applications. Obtaining such genome-wide location maps directly is often invasive and resource-intensive, so it is common to impute binding locations from DNA sequence or measures of chromatin accessibility. We introduce DeepATAC, a deep-learning approach for imputing binding locations that uses both DNA sequence and chromatin accessibility as measured by ATAC-seq. DeepATAC significantly outperforms current approaches such as FIMO motif predictions overlapped with ATAC-seq peaks, and models based only on DNA sequence, such as DeepSEA. Visualizing the input importances for the DeepATAC model reveals DNA sequence motifs and ATAC-seq signal patterns that are important for predicting binding events. The R package implementation and analysis pipeline of DeepATAC are available at https://github.com/hiranuma/deepatac.

DeepBound: Accurate Identification of Transcript Boundaries via Deep Convolutional Neural Fields

Mingfu Shao\textsuperscript{1,1,*}, Jianzhu Ma\textsuperscript{2,1} and Sheng Wang\textsuperscript{3,1,*}

\textsuperscript{1} Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, and
\textsuperscript{2} School of Medicine, University of California San Diego, La Jolla, CA 92030, and
\textsuperscript{3} Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.

\textsuperscript{*} The three authors contribute equally to this work.
\textsuperscript{1} To whom correspondence should be addressed.

Associate Editor: XXXXXX

Abstract

Motivation: Reconstructing the full-length expressed transcripts (\textit{a.k.a.} the transcript assembly problem) from the short sequencing reads produced by RNA-seq protocol plays a central role in identifying novel genes and transcripts as well as in studying gene expressions and gene functions. A crucial step in transcript assembly is to accurately determine the splicing junctions and boundaries of the expressed transcripts from the reads alignment. In contrast to the splicing junctions that can be efficiently detected from spliced reads, the problem of identifying boundaries remains open and challenging, due to the fact that the signal related to boundaries is noisy and weak.

Results: We present DeepBound, an effective approach to identify boundaries of expressed transcripts from RNA-seq reads alignment. In its core DeepBound employs deep convolutional neural fields to learn the hidden distributions and patterns of boundaries. To accurately model the transition probabilities and to solve the label-imbalance problem, we newly incorporate the AUC (area under the curve) score into the optimizing objective function. To address the issue that deep probabilistic graphical models requires large number of labeled training samples, we propose to use simulated RNA-seq datasets to train our model. Through extensive experimental studies on both simulation datasets of two species and biological datasets, we show that DeepBound consistently and significantly outperforms the two existing methods.
Using Neural Networks to Improve Single Cell RNA-Seq Data Analysis

Chieh Lin\textsuperscript{1}, Siddhartha Jain\textsuperscript{2}, Hannah Kim\textsuperscript{3}, Ziv Bar-Joseph\textsuperscript{1,3,*}
\textsuperscript{1}Machine Learning Department, \textsuperscript{2}Computer Science Department, \textsuperscript{3}Computational Biology Department
School of Computer Science, Carnegie Mellon University
* Corresponding author, zibj@cs.cmu.edu

Abstract

While only recently developed, the ability to profile expression data in single cells (scRNA-Seq) has already led to several important studies and findings. However, this technology has also raised several new computational challenges including questions related to handling the noisy and sometimes incomplete data, how to identify unique group of cells in such experiments and how to determine the state or function of specific cells based on their expression profile. To address these issues we develop and test a method based on neural networks (NN) for the analysis and retrieval of single cell RNA-Seq data. We tested various NN architectures, some biologically motivated, and used these to obtain a reduced dimension representation of the single cell expression data. We show that the NN method improves upon prior methods in both, the ability to correctly group cells in experiments not used in the training and the ability to correctly infer cell type or state by querying a database of tens of thousands of single cell profiles. Such database queries (which can be performed using our web server) will enable researchers to better characterize cells when analyzing heterogeneous scRNA-Seq samples.