Deep Learning for Computational Biology

CS260

January 11, 2018

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 it 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

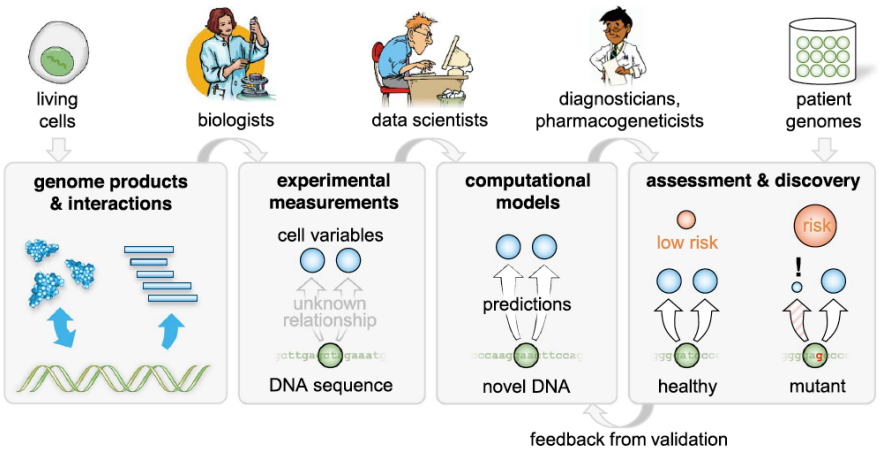
- The final grade will be based on paper presentations and participation in the class
- While the presenter is responsible to describe the paper in detail, it would be beneficial for the others to review the paper in advance to be able to participate in the discussion
- I will have a selection of papers, 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



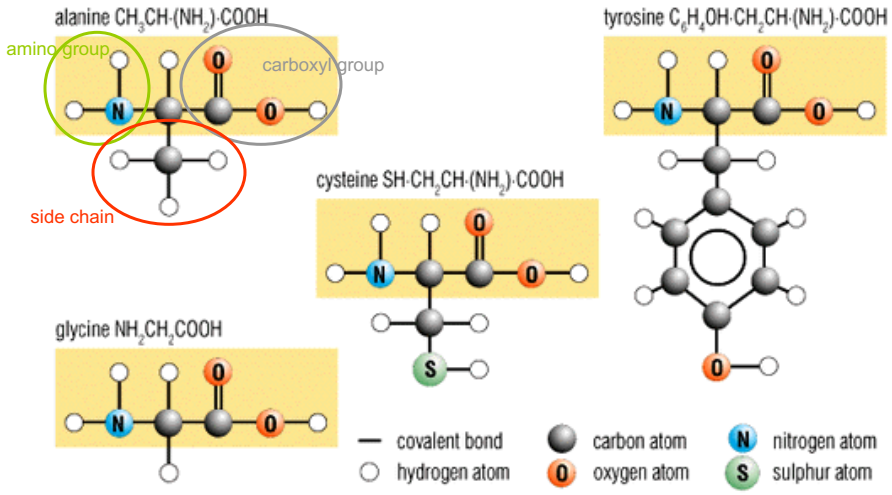
Proteins

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_α), 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



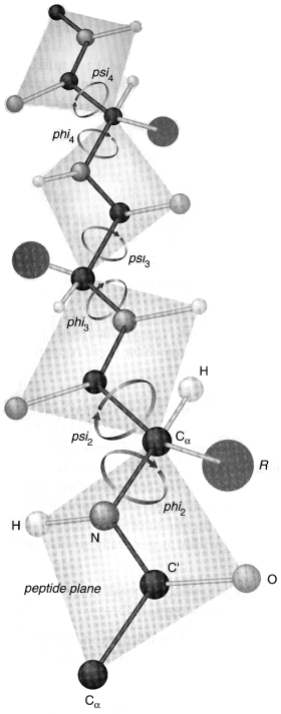
Proteins

- 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

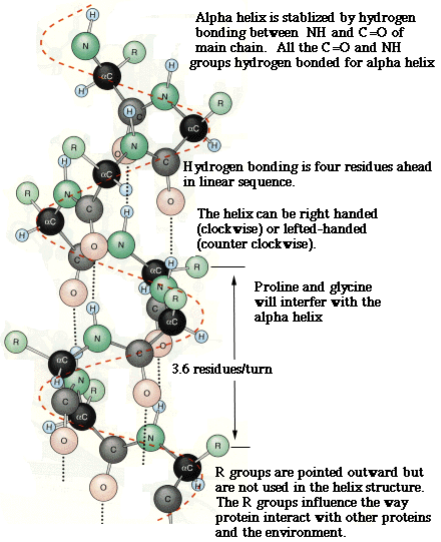
- *Primary* structure: the linear sequence of amino acids, ordered from the N-terminal (amino group) to C-terminal (carboxy group)
- *Secondary* structure: α-helices and β-sheets
- *Tertiary* structure: the 3D conformation (*folding*) in space

- 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



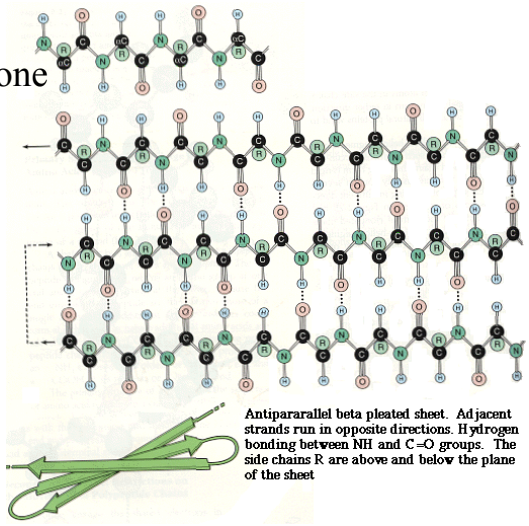
Alpha helix

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



Beta sheet

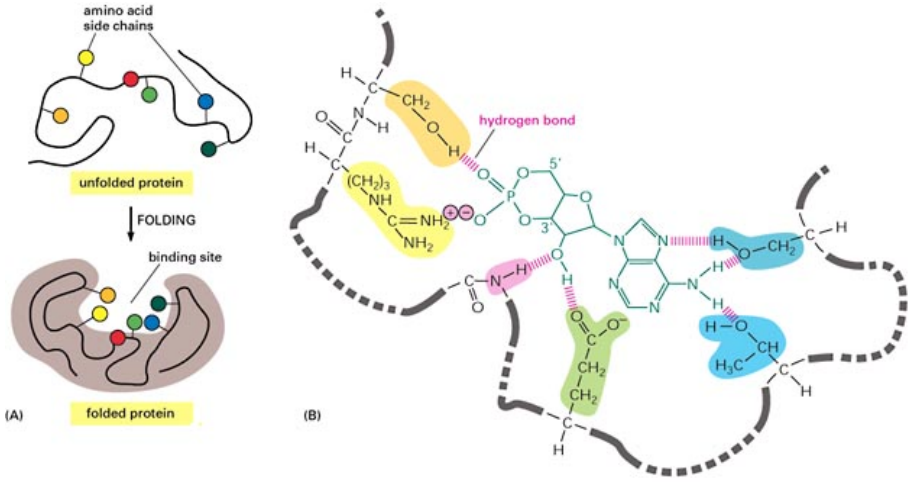
- Regions of extended (nearly linear) backbone conformation with phi≈135 and psi≈135
- Hydrogen bonds
- Two types
  - Parallel
  - Anti-parallel



Antiparallel beta pleated sheet. Adjacent strands run in opposite directions. Hydrogen bonding between NH and C=O groups. The side chains R are above and below the plane of the sheet

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)

SCIENTIFIC REPORTS

OPEN

Received: 28 June 2015
Accepted: 26 November 2015
Published: 11 January 2016

Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields

Sheng Wang[1,2], Jian Peng[3], Jianzhu Ma[1] & Jinbo Xu[1]

Protein secondary structure (SS) prediction is important for studying protein structure and function. When only the sequence (profile) information is used as input feature, currently the best predictors can obtain ~80% Q3 accuracy, which has not been improved in the past decade. Here we present DeepCNF (Deep Convolutional Neural Fields) for protein SS prediction. DeepCNF is a Deep Learning extension of Conditional Neural Fields (CNF), which is an integration of Conditional Random Fields (CRF) and shallow neural networks. DeepCNF can model not only complex sequence-structure relationship by a deep hierarchical architecture, but also interdependency between adjacent SS labels, so it is much more powerful than CNF. Experimental results show that DeepCNF can obtain ~84% Q3 accuracy, ~85% SOV score, and ~72% Q8 accuracy, respectively, on the CASP and CAMEO test proteins, greatly outperforming currently popular predictors. As a general framework, DeepCNF can be used to predict other protein structure properties such as contact number, disorder regions, and solvent accessibility.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

Sheng Wang[☯], **Siqi Sun**[☯], **Zhen Li**, **Renyu Zhang**, **Jinbo Xu***

Toyota Technological Institute at Chicago, Chicago, Illinois, United States of America

☯ These authors contributed equally to this work.
* jinboxu@gmail.com

OPEN ACCESS

Citation: Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol 13(1): e1005324. doi:10.1371/journal.pcbi.1005324

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: September 14, 2016

Accepted: December 20, 2016

Published: January 5, 2017

Data Availability Statement: 1) The PDB25 list is available at http://dunbrack.fccc.edu/PISCES.php. 2) The CASP11 test proteins are available at the CASP web site (http://predictioncenter.org/). 3) The other data lists are provided in the paper and the Supporting Information files.

Funding: This work is supported by National Institutes of Health grant R01GM089753 to JX and National Science Foundation grant DBI-1564955 to JX. The authors are also grateful to the support of Nvidia Inc. and the computational resources provided by XSEDE through the grant MCB150134

Abstract

Motivation

Protein contacts contain key information for the understanding of protein structure and function and thus, contact prediction from sequence is an important problem. Recently exciting progress has been made on this problem, but the predicted contacts for proteins without many sequence homologs is still of low quality and not very useful for de novo structure prediction.

Method

This paper presents a new deep learning method that predicts contacts by integrating both evolutionary coupling (EC) and sequence conservation information through an ultra-deep neural network formed by two deep residual neural networks. The first residual network conducts a series of 1-dimensional convolutional transformation of sequential features; the second residual network conducts a series of 2-dimensional convolutional transformation of pairwise information including output of the first residual network, EC information and pairwise potential. By using very deep residual networks, we can accurately model contact occurrence patterns and complex sequence-structure relationship and thus, obtain higher-quality contact prediction regardless of how many sequence homologs are available for proteins in question.

Results

Our method greatly outperforms existing methods and leads to much more accurate contact-assisted folding. Tested on 105 CASP11 targets, 76 past CAMEO hard targets, and 398 membrane proteins, the average top L long-range prediction accuracy obtained by our method, one representative EC method CCMpred and the CASP11 winner MetaPSICOV is 0.47, 0.21 and 0.30, respectively; the average top L/10 long-range accuracy of our method, CCMpred and MetaPSICOV is 0.77, 0.47 and 0.59, respectively. Ab initio folding using our predicted contacts as restraints but without any force fields can yield correct folds (i.e., TMscore>0.6) for 203 of the 579 test proteins, while that using MetaPSICOV- and CCMpred-predicted contacts can do so for only 79 and 62 of them, respectively. Our contact-assisted models also have much better quality than template-based models especially for membrane proteins. The 3D models built

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 28 no. 19 2012, pages 2449–2457
doi:10.1093/bioinformatics/bts475

Structural bioinformatics

Advance Access publication July 30, 2012

Deep architectures for protein contact map prediction

Pietro Di Lena[1,2], Ken Nagata[1,2] and Pierre Baldi[1,2,*]

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

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Residue–residue contact prediction is important for protein structure prediction and other applications. However, the accuracy of current contact predictors often 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 contact map prediction using three steps of increasing resolution. First, we use 2D recursive neural networks to predict coarse contacts and orientations between 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 over both space and time. We train the architecture on a large set of non-redundant proteins and test it on a large set of non-homologous domains, as well as on the set of protein domains used for contact prediction in the two most recent CASP8 and CASP9 experiments. For long-range contacts, the accuracy of the new CMAPpro predictor is close to 30%, a significant increase over existing approaches.

routinely reported at CASP for the best predictors (Ezkurdia *et al.*, 2009; Kryshtafovych *et al.*, 2011), suggests that contact prediction is not yet accurate enough to be systematically useful for *ab initio* protein structure prediction or engineering.

In broad terms, there are four main approaches for residue–residue contact prediction. Machine learning approaches use methods such as neural networks (Fariselli *et al.*, 2001; Punta and Rost, 2005; Shackelford and Karplus, 2007), recursive neural networks (Baldi and Pollastri, 2003; Vullo *et al.*, 2006), support vector machines (Cheng and Baldi, 2007) and hidden Markov models (Björkholm *et al.*, 2009) to learn how to predict contact probabilities from a training set of experimentally determined protein structures. Inputs to these approaches typically include predicted secondary structure, predicted solvent accessibility as well as evolutionary information in the form of profiles. Template-based approaches use homology or threading methods to identify structurally similar templates from which residue–residue contacts are then inferred (Misura *et al.*, 2006; Skolnick *et al.*, 2004). Correlated mutations approaches apply statistical measures, such as Pearson correlation (Göbel *et al.*, 1994; Olmea and Valencia, 1997) and mutual information (Burger and van

Cao et al. BMC Bioinformatics
DOI 10.1186/s12859-016-1405-y

BMC Bioinformatics

RESEARCH ARTICLE Open Access

DeepQA: improving the estimation of single protein model quality with deep belief networks

Renzhi Cao[1], Debswapna Bhattacharya[2], Jie Hou[3] and Jianlin Cheng[3,4*]

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, physio-chemical 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.rnet.missouri.edu/DeepQA/.

Bioinformatics, 2016, 1–8
doi: 10.1093/bioinformatics/btw678
Original Paper

Structural bioinformatics

Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks

Jack Hanson[1,*], Yuedong Yang[2,*], Kuldip Paliwal[1] and Yaoqi Zhou[2,*]

[1]Signal Processing Laboratory, Griffith University, Brisbane 4122, Australia and [2]Institute for Glycomics, Griffith University, Gold Coast 4215, Australia

*To whom correspondence should be addressed.
Associate editor: Anna Tramontano

Received on June 7, 2016; revised on September 29, 2016; editorial decision on October 22, 2016; accepted on October 26, 2016

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 MobiDB, 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.

Nucleic Acids Research, 2016 1
doi: 10.1093/nar/gkw306

RaptorX-Property: a web server for protein structure property prediction

Sheng Wang[1,2,*,†], Wei Li[3,†], Shiwang Liu[3] and Jinbo Xu[1,*]

[1]Toyota Technological Institute at Chicago, Chicago, IL, USA, [2]Department of Human Genetics, University of Chicago, Chicago, IL, USA and [3]School 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% Q8 accuracy for 8-state SS, ∼66% Q3 accuracy for 3-state solvent accessibility, and ∼0.89 area under the ROC curve (AUC) for disorder prediction.

tural properties from amino acid sequence alone, without using any template information (7).

However, the prediction accuracy of protein structural properties, while without exploiting experimentally-solved structures (i.e. templates), is still far away from satisfactory. Taking 3-state secondary structure prediction as an example, when template information is not used and only sequence profile is considered, so far the best Q3 accuracy is ∼80% obtained by a few predictors such as PSIPRED (8) and JPRED (7), which is significantly lower than the estimated prediction accuracy limit 88–90% (9). Such a gap motivates us to develop a better method to further improve SS prediction. A similar trend is observed on solvent accessibility prediction with three-state accuracy ∼60% obtained by SPINE-X (10) and SANN (11). To further increase prediction accuracy, we will need a more sophisticated method that can model the complex sequence-structure relationship in a much better way.

This paper presents RaptorX Property, a web server predicting protein structure property solely based on protein sequence or sequence profile. A profile is derived from multiple sequence alignment (MSA) of sequence homologs in a protein family (12). To predict structure properties, this server employs a new machine learning model DeepCNF

Deep learning with feature embedding for compound-protein interaction prediction

Fangping Wan
wfp15@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

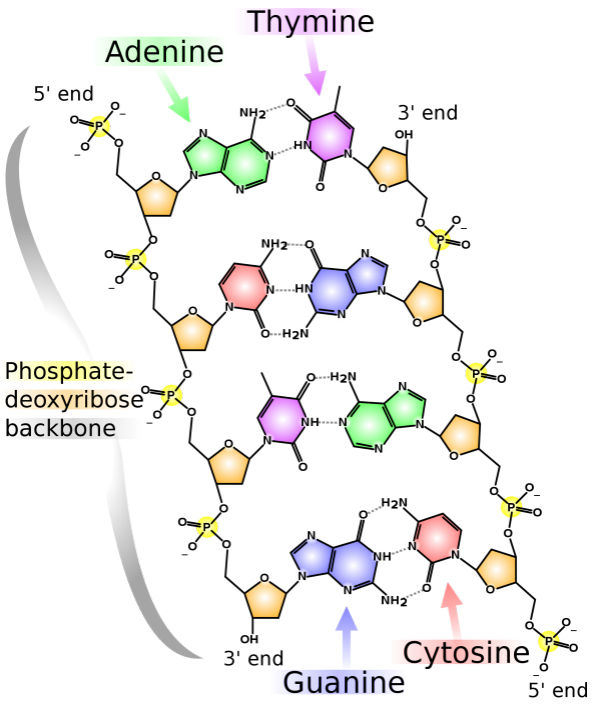
DNA Binding
RNA Binding
Enhancers/Promoters
Alternative Splicing
MicroRNA post-transcriptional regulation

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

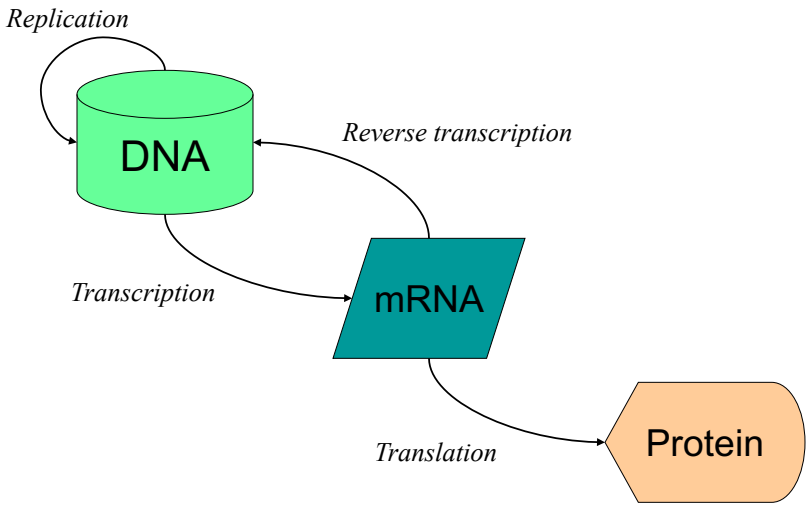


RNA

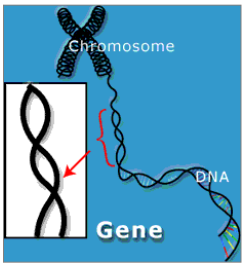
- Single stranded
- Uracil (U) instead of thymine (T)
- Different types of RNA
  - mRNA (messenger RNA)
  - tRNA (transfer RNA)
  - rRNA (ribosomal RNA)

  … 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.

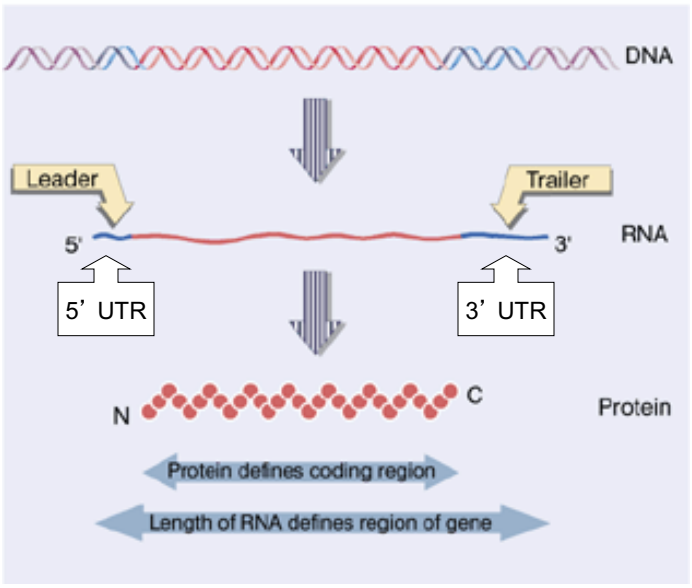
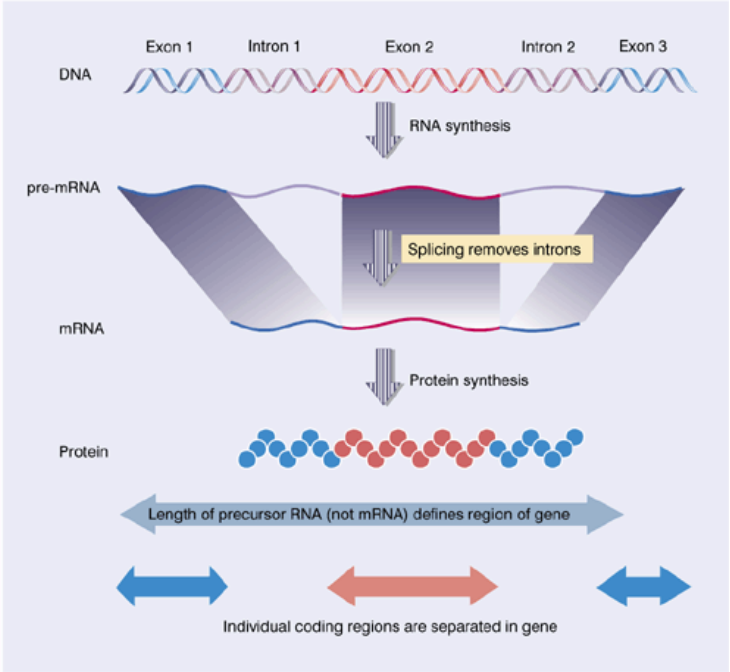
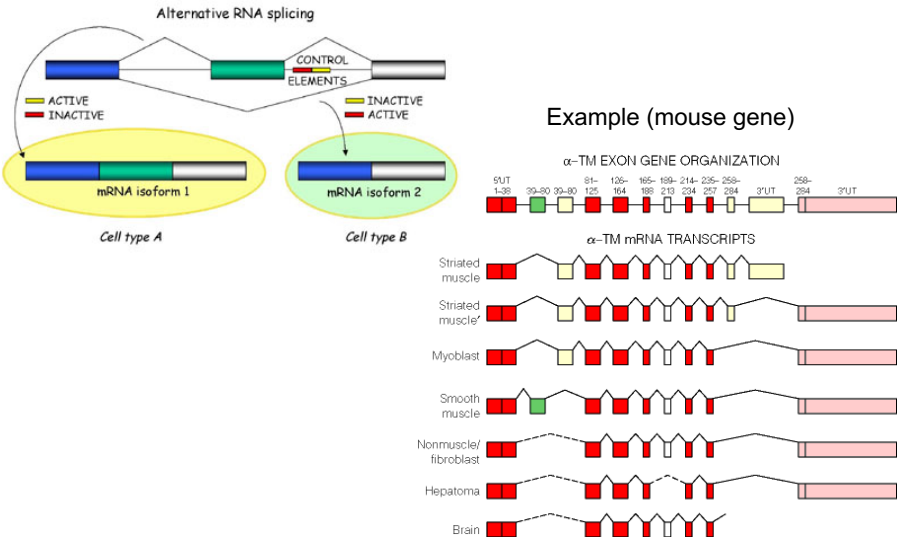


Figure 2.10 Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.



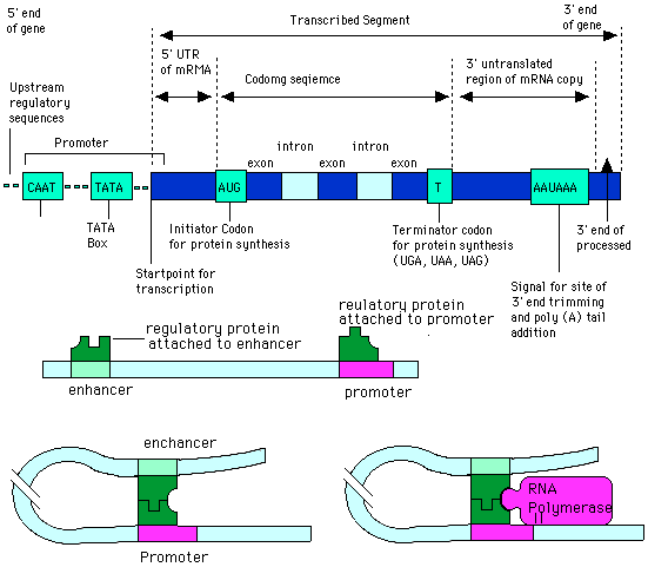
Alternative splicing



Transcription

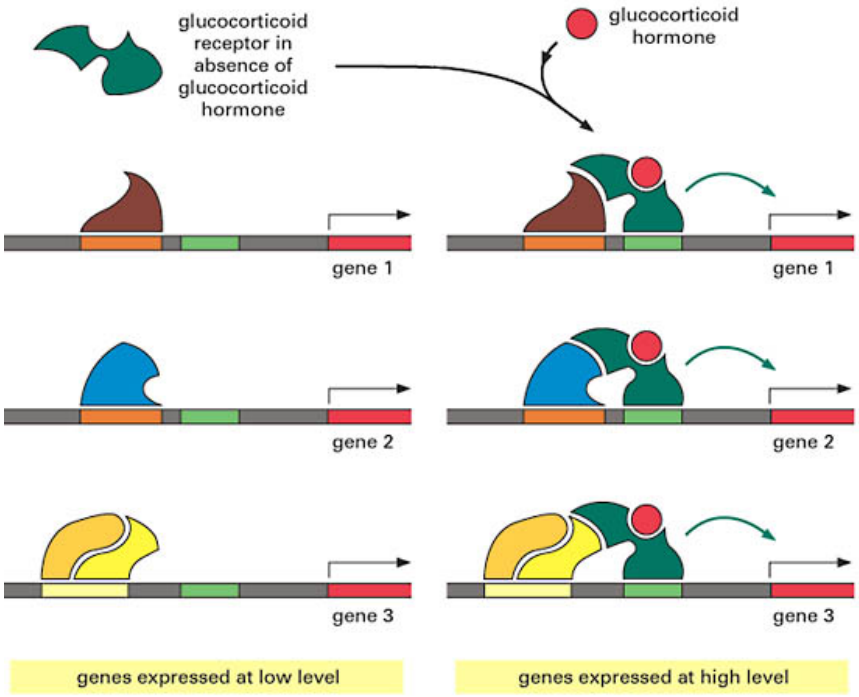
- *Promoter:* a region of DNA involved in binding of RNA polymerase to initiate transcription
- *Enhancer:* a region of DNA that increases the utilization of (some) promoters (it can function in either orientations and any location relative to the promoter)
- *Repressor:* a region of DNA that decreases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

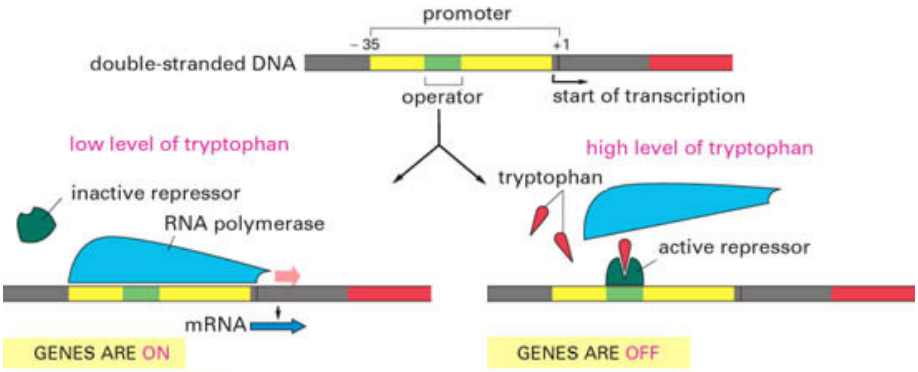
Promoters and Enhancers

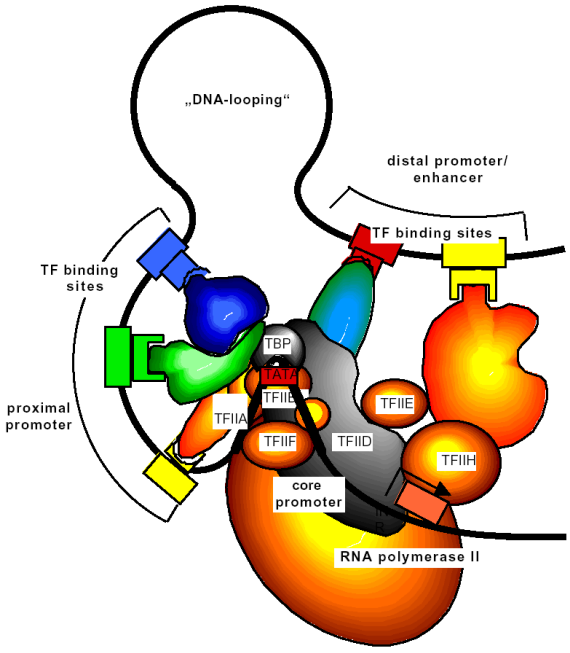


Transcription control

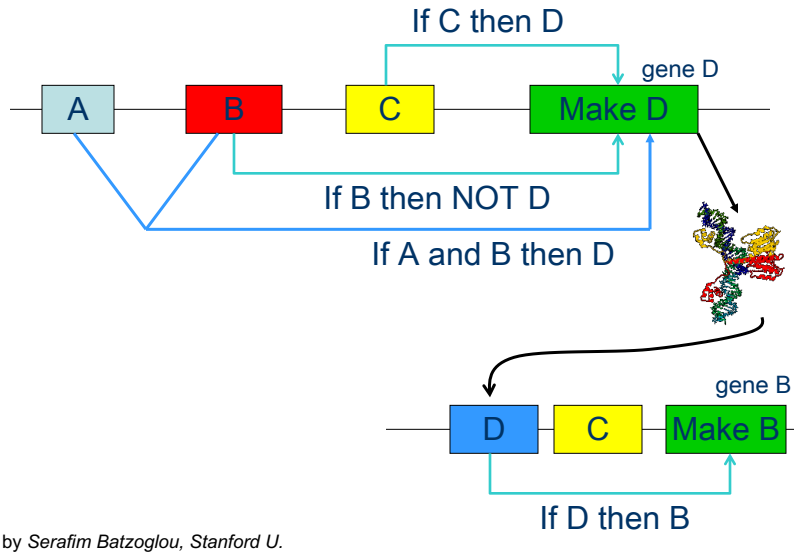
- Different factors are involved in the transcription machinery
  - binding of transcription factors to DNA
  - ability of DNA to bend
  - relative location of the binding sites
  - interaction between transcription factors
  - DNA methylation, nucleosomes (epigenetics)
  - presence C*p*G islands ("*p*" is for phosphate)
  - …


glucocorticoid receptor in absence of glucocorticoid hormone
glucocorticoid hormone
gene 1
gene 1
gene 2
gene 2
gene 3
gene 3
genes expressed at low level
genes expressed at high level


promoter
– 35
+1
double-stranded DNA
operator
start of transcription
low level of tryptophan
high level of tryptophan
tryptophan
inactive repressor
RNA polymerase
active repressor
mRNA
GENES ARE ON
GENES ARE OFF

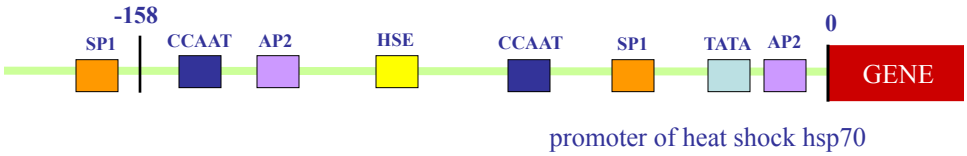


Genetic “circuits”



Slide by *Serafim Batzoglou, Stanford U.*

Example: A Human heat shock protein



promoter of heat shock hsp70

- 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.

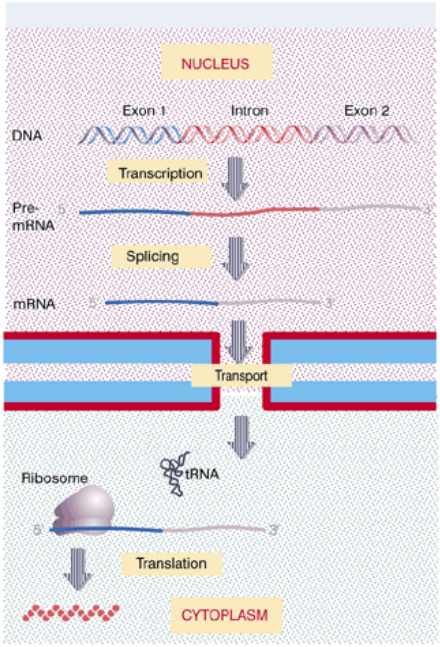
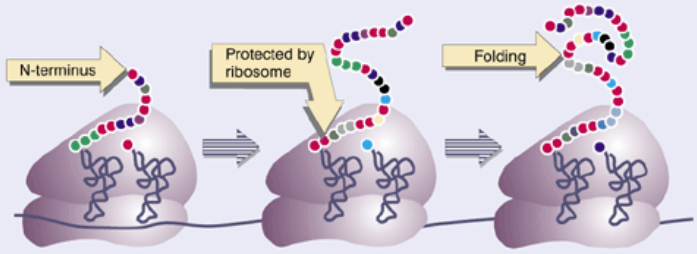
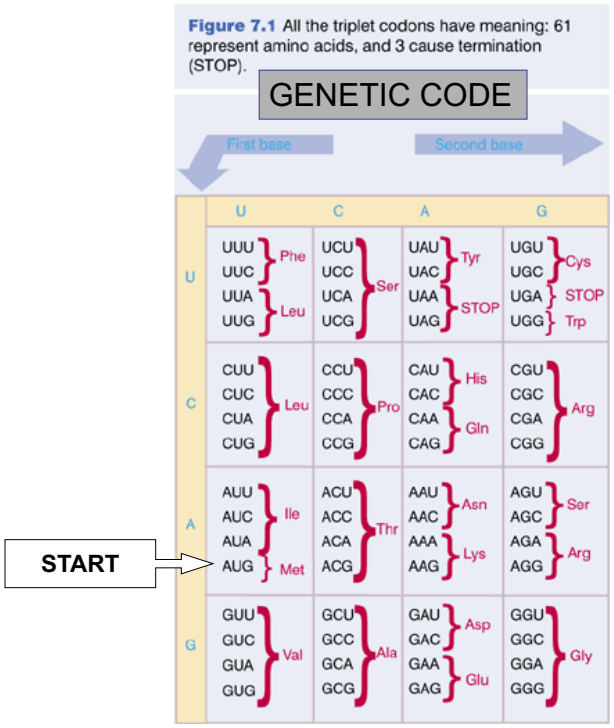


Figure 5.8 A polyribosome consists of an mRNA being translated simultaneously by several ribosomes moving in the direction from 5' to 3'. Each ribosome has two tRNA molecules: one carrying the nascent protein, the second carrying the next amino acid to be added.





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

Bioinformatics, 32, 2016, i121–i127
doi: 10.1093/bioinformatics/btw255
ISMB 2016

Convolutional neural network architectures for predicting DNA–protein binding

Haoyang Zeng, Matthew D. Edwards, Ge Liu and David K. Gifford*

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Convolutional neural networks (CNN) have outperformed conventional methods in modeling the sequence specificity of DNA–protein binding. Yet inappropriate CNN architectures can yield poorer performance than simpler models. Thus an in-depth understanding of how to match CNN architecture to a given task is needed to fully harness the power of CNNs for computational biology applications.

Results: We present a systematic exploration of CNN architectures for predicting DNA sequence binding using a large compendium of transcription factor datasets. We identify the best-performing architectures by varying CNN width, depth and pooling designs. We find that adding convolutional kernels to a network is important for motif-based tasks. We show the benefits of CNNs in learning rich higher-order sequence features, such as secondary motifs and local sequence context, by comparing network performance on multiple modeling tasks ranging in difficulty. We also demonstrate how careful construction of sequence benchmark datasets, using approaches that control potentially confounding effects like positional or motif strength bias, is critical in making fair comparisons between competing methods. We explore how to establish the sufficiency of training data for these learning tasks, and we have created a flexible cloud-based framework that permits the rapid exploration of alternative neural network architectures for problems in computational biology.

Published in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins

Hamid Reza Hassanzadeh
Department of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
Email: hassanzadeh@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 *cis*-regulatory specific sub-regions 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 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 reckons with the contributions made by individual sub-regions 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-vitro 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.**

molecular techniques to pinpoint the locations on DNA where these factors bind, which in itself requires the ability to measure precisely the binding affinity between these molecules. With the advances in high-throughput technologies in the past decade several in-vivo [9], [18] and in-vitro [2], [10], [12] techniques have been invented and upgraded to address this important and yet challenging task. Unfortunately, none of these methods are able to generate results that are interpretable by biologists, but instead each generates a large volume of noisy, erroneous and low-resolution measurements for tens of thousands sequence probes. As a result, the outcome of such experiments need to be processed through downstream analysis pipelines to elicit useful information. In this study, we use data from Protein Binding Microarrays (PBM) experiments to evaluate our proposed method. PBM [2] is a recent in-vitro high-throughput technology that can massively measure relative binding preferences of DNA probes for a given transcription factor. The binding preference in a PBM experiment is directly related to the measured spot intensities which are scanned and recorded for later analysis. We apply our method to the data produced by a set of PBM experiments and try to predict the binding preferences for the test probes.

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

Shashank Singh[1], Yang Yang[2], Barnabás Póczos[1], and Jian Ma[2,*]

[1]Machine Learning Department
[2]Computational 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 Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

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 or as a 'mutation map' that indicates how variations affect binding within a specific sequence.

come in qualitatively different forms. Protein binding microarrays (PBMs)[8] and RNAcompete assays[9] provide a specificity coefficient for each probe sequence, whereas chromatin immunoprecipitation (ChIP)-seq[10] provides a ranked list of putatively bound sequences of varying length, and HT-SELEX[11] generates a set of very high affinity sequences. Second, the quantity of data is large. A typical high-throughput experiment measures between 10,000 and 100,000 sequences, and it is computationally demanding to incorporate them all. Third, each data acquisition technology has its own artifacts, biases and limitations, and we must discover the pertinent specificities despite these unwanted effects. For example, ChIP-seq reads often localize to "hyper-ChIPable" regions of the genome near highly expressed genes[12].

DeepBind (**Fig. 1**) addresses the above challenges. (i) It can be applied to both microarray and sequencing data; (ii) it can learn from millions of sequences through parallel implementation on a graphics processing unit (GPU); (iii) it generalizes well across technologies, even without correcting for technology-specific biases; (iv) it can

Published online 13 October 2015

Nucleic Acids Research, 2016, Vol. 44, No. 4 e32
doi: 10.1093/nar/gkv1025

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

Sai Zhang[1], Jingtian Zhou[2,†], Hailin Hu[2,†], Haipeng Gong[3,4], Ligong Chen[2], Chao Cheng[5,*] and Jianyang Zeng[1,4,*]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China, [2]Department of Pharmacology and Pharmaceutical Sciences, School of Medicine, Tsinghua University, Beijing 100084, China, [3]School of Life Sciences, Tsinghua University, Beijing 100084, China, [4]MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China and [5]Department of Genetics, Institute for Quantitative Biomedical Sciences, Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA

Received April 13, 2015; Revised September 11, 2015; Accepted September 28, 2015

ABSTRACT

RNA-binding proteins (RBPs) play important roles in the post-transcriptional control of RNAs. Identifying RBP binding sites and characterizing RBP binding preferences are key steps toward understanding the basic mechanisms of the post-transcriptional gene regulation. Though numerous computational methods have been developed for modeling RBP binding preferences, discovering a complete structural representation of the RBP targets by integrating their available structural 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, which can be further used to predict novel candidate binding sites and discover potential binding motifs. Through testing on the real `CLIP-seq` datasets, 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 conducted 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 polypyrimidine tract-binding protein (PTB), which also provides a new evidence to support the view that RBPs may own specific tertiary structural binding preferences. In particular, the tests on the internal ribosome entry site (IRES) segments yield satisfiable results with experimental support from the literature and further demonstrate the necessity of incorporating RNA tertiary structural information into the prediction model. The source code of our approach can be found in https://github.com/thucombio/deepnet-rbp.

INTRODUCTION

RNA-binding proteins (RBPs) play important roles in various cellular processes, such as alternative splicing, RNA editing, mRNA localization and translational regulation (1). RBPs contain several special RNA-binding domains (RBDs), e.g. the RNA recognition motif (RRM) and the hnRNP K-homology (KH) domains, which recognize their target sites related to the RNA primary sequence and the corresponding structural profiles (2). Although it has been shown that several important diseases, such as neurodegenerative disorders, cancers and cardiovascular diseases, can be caused by the dysfunctions of certain RBPs (3,4), relatively few RBPs have been well characterized. Therefore, identifying RNA–protein interactions and modeling RBP binding preferences are important for decoding the post-transcriptional processes involving RBPs and their mechanisms of pathogenesis in human diseases.

Recently, the advent of high-throughput experimental methods, such as the cross-linking immunoprecipitation coupled with high-throughput sequencing (`CLIP-seq`) protocols, has greatly advanced the genome-wide studies of RNA–protein interactions (5–8). Despite the success stories 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

Shuya Li[1,†], Fanghong Dong[2,†], Yuexin Wu[2,3,†], Sai Zhang[2], Chen Zhang[2], Xiao Liu[1], Tao Jiang[4,5], and Jianyang Zeng[2,*]

[1]School of Life Sciences, Tsinghua University, Beijing, China.
[2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
[3]Present address: Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
[4]Department of Computer Science and Engineering, University of California, Riverside, CA.
[5]MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST / Department of Computer Science and Technology, Tsinghua University, Beijing, China.
*To whom correspondence should be addressed. Email: zengjy321@tsinghua.edu.cn.
†These authors contributed equally to this work.

Abstract

Characterizing the binding behaviors of RNA-binding proteins (RBPs) is important for understanding their functional roles in gene expression regulation. However, current high-throughput experimental methods for identifying RBP targets, 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 DeBooster, 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 DeBooster 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 DeBooster in understanding the regulatory func-

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.

BIOINFORMATICS

Vol. 30 ISMB 2014, pages i121–i129
doi:10.1093/bioinformatics/btu277

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.

Bioinformatics, 32(12), 2016, 1832–1839
doi: 10.1093/bioinformatics/btw074
Advance Access Publication Date: 11 February 2016
Original Paper

Gene expression

Gene expression inference with deep learning

Yifei Chen[1,4,†], Yi Li[1,†], Rajiv Narayan[2], Aravind Subramanian[2] and Xiaohui Xie[1,3,*]

[1]Department of Computer Science, University of California, Irvine, CA 92697, USA, [2]Broad Institute of MIT And Harvard, Cambridge, MA 02142, USA, [3]Center for Complex Biological Systems, University of California, Irvine, CA 92697, USA and [4]Baidu Research-Big Data Lab, Beijing, 100085, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Inanc Birol

Received on August 5, 2015; revised on December 14, 2015; accepted on February 3, 2016

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 compendium of expression profiling over thousands of samples is still very expensive. Recognizing that gene expressions are often highly correlated, researchers from the NIH LINCS 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 LINCS program is currently based on linear regression (LR), limiting its accuracy since it does not capture complex nonlinear relationship between expressions of genes.
Results: We present a deep learning method (abbreviated as D-GEX) to infer the expression of target genes from the expression of landmark genes. We used the microarray-based Gene Expression Omnibus dataset, consisting of 111K 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 99.97% of the target genes. We also tested the performance of our learned model on an independent RNA-Seq-based GTEx dataset, which consists of 2921 expression profiles. Deep learning still outperforms LR with 6.57% relative improvement, and achieves lower error in 81.31% of the target genes.

Bioinformatics, 32, 2016, i8–i17
doi: 10.1093/bioinformatics/btw243
ISMB 2016

What time is it? Deep learning approaches for circadian rhythms

Forest Agostinelli[1,*], Nicholas Ceglia[1], Babak Shahbaba[2], Paolo Sassone-Corsi[3] and Pierre Baldi[1,3,*]

[1]Department of Computer Science, [2]Department of Statistics and [3]Department 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 fashion 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 h, using 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 Li[1,2], Wenqiang Shi[1], and Wyeth W. Wasserman*[1]

[1]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

[2]Information 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 DECRES, 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 DECRES, 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).

Pan *et al. BMC Genomics* (2016) 17:582
DOI 10.1186/s12864-016-2931-8

BMC Genomics

METHODOLOGY ARTICLE Open Access

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

Xiaoyong Pan[1,4†], Yong-Xian Fan[2†], Junchi Yan[3] and Hong-Bin Shen[1*]

Abstract

Background: Non-coding RNAs (ncRNAs) play crucial roles in many biological processes, such as post-transcription of gene regulation. ncRNAs mainly function through interaction with RNA binding proteins (RBPs). To understand the function of a ncRNA, a fundamental step is to identify which protein is involved into 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 lncRNA-protein interaction dataset with an accuracy of 0.891, sensitivity of 0.939, specificity of 0.831, precision of 0.945 and Matthews correlation coefficient of 0.784, 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 20 % on some tested benchmarked datasets. In addition, we further apply IPMiner for large-scale prediction of ncRNA-protein network, that achieves promising prediction performance.

Conclusion: By integrating deep neural network and stacked ensembling, from simple sequence composition features, IPMiner can automatically learn high-level abstraction features, which had strong discriminant ability for RNA-protein detection. IPMiner achieved high performance on our constructed lncRNA-protein benchmark dataset and other RNA-protein datasets. IPMiner tool is available at http://www.csbio.sjtu.edu.cn/bioinf/IPMiner.

Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters

Yifeng Li, Chih-Yu Chen, and Wyeth W. Wasserman(✉)

Centre for Molecular Medicine and Therapeutics, University of British Columbia,
950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada
{yifeng,juliec,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 structures 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 5 November 2014

Nucleic Acids Research, 2015, Vol. 43, No. 1 e6
doi: 10.1093/nar/gku1058

DEEP: a general computational framework for predicting enhancers

Dimitrios Kleftogiannis[1], Panos Kalnis[1] and Vladimir B. Bajic[2,*]

[1]Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia and [2]Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

Received June 26, 2014; Revised October 04, 2014; Accepted October 16, 2014

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 bioinformatics research. Although existing methodologies increased the number of computationally predicted enhancers, performance inconsistency of computational models across different cell-lines, class imbalance within the learning sets and *ad hoc* rules for selecting enhancer candidates for supervised learning, are some key questions that require further examination. In this study we developed DEEP, a novel ensemble prediction framework. DEEP integrates three components with diverse characteristics that streamline the analysis of enhancer's properties in a great variety of cellular conditions. In our method we train many individual classification models that we combine to classify DNA regions as enhancers or non-enhancers. DEEP uses features derived from histone modification marks or attributes coming from sequence characteristics. Experimental results indicate that DEEP performs better than four state-of-the-art methods on the ENCODE data. We report the first computational enhancer prediction results on FANTOM5 data where DEEP achieves 90.2% accuracy and 90% geometric mean (GM) of specificity and sensitivity across 36 different tissues. We further present results derived using *in vivo*-derived enhancer data from VISTA database. DEEP-VISTA, when tested on an independent test set, achieved GM of 80.1% and accuracy of 89.64%. DEEP framework is publicly available at http://cbrc.kaust.edu.sa/deep/.

INTRODUCTION

Transcription regulation in human genes is a complex process (1,2). Promoters are cis-regulatory regions, which serve as anchor points for recruiting multiprotein complexes required for transcription. Although these regions have been extensively studied, their underlying transcriptional mechanism is not yet fully understood (3). Recent advances in high-throughput experiments like the 3C technology indicate that interactions between proximal and distal regulatory elements orchestrate gene expression between different cell types. In contrast to proximal elements, distal elements are not located near to the genes whose activity they affect, and can be located 20 kb or further away, or even can be located at different chromosomes. In addition, their functional mechanism appears to be independent of the upstream/downstream location of the genes they target. The better-characterized distal regulatory elements in eukaryotes are enhancers, silencers and insulators (4,5). Providing an accurate definition for these regulatory elements is not an easy task since they may have different roles depending on the cellular state (i.e. can be active or inactive, or can assume non-enhancer function) and their functional mechanism is not yet fully known. In the line with (6) we characterize enhancers as *cis*-acting DNA regulatory elements that increase the transcriptional output of the distal target genes. Enhancers activate gene transcription by recruiting transcription factors (TFs) and their complexes. For this reason, enhancer regions frequently contain clusters of binding sites of various TFs that vary across different cells and tissues. On the other hand, silencers, repressors and insulators have practically negative effects on the cellular transcriptional output either through recruitment of transcriptional repressor proteins (7), or by preventing the spread of heterochromatin (8).

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

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

Seunghyun Park
Electrical and Computer Eng.
Seoul National University
Seoul 08826, Korea
Electrical Engineering
Korea University
Seoul 02841, 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-sequence 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 deepTarget delivers a quantum leap in the long-standing challenge of robust miRNA target prediction. [availability: http://data.snu.ac.kr/pub/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, miRNAs are derived from the precursor miRNAs (pre-miRNAs) and then exhibit their regulatory function by binding to the target sites present in mRNAs. Two types of computational problems about miRNAs thus naturally arise in bioinformatics: miRNA host 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 30022, 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 sizes of gene neighborhoods, but that 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 20Kb gene neighborhoods labeled by binarized levels of target gene expression. After supervised training of the model we achieve 77% accuracy as estimated by 10-fold CV.

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

and mouse ENCODE projects(ENCODE Project Consortium, 2011; Mouse ENCODE Consortium et al., 2012) for the generation of hundreds of assays targeting specific transcription factors (TFs) on a variety of cell lines. ChIP-Seq derived TFos 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, but 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, 1965; 1968), 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., 2004) 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

PROCEEDINGS Open Access

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 (APBC 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 it 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 models 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]

[1]Computer Science, Stanford University, Stanford, 94305, USA
[2]Biomedical Informatics, Stanford University, Stanford, 94305, USA
[3]Genetics, 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,†], Hailin Hu [2,†], Tao Jiang [3,4,5], Lei Zhang [2,*] and Jianyang Zeng [1,*]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China.
[2] School of Medicine, Tsinghua University, Beijing 100084, China.
[3] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA.
[4] MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
[5] Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA.
[†] These authors contributed equally to this work.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received 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., GTI-seq and QTI-seq, provides abundant data for systematically studying the general principles of translation initiation and the development of computational method 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 QTI-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[2,3,4], and Jianyang Zeng[1,*]

[1] Institute for Interdisciplinary Information Science, Tsinghua University, Beijing 100084, China.
[2] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA.
[3] MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
[4] Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA.
[*] To whom correspondence should be addressed.
[*] E-mail: zengjy321@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, snRNAs, snoRNAs and mRNAs. 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 pseudourdines (especially in mRNAs) still remains largely unclear. Though thousands of RNA pseudouridylation 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 (PseudoUridyLation Sites Estimator), 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 outperform conventional learning models and achieve high prediction accuracy, thus enabling us to further characterize the transcriptome-wide landscape of pseudouridine 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

Naozumi Hiranuma[1], Scott Lundberg[1], Su-In Lee[1,2]

[1] Paul G. Allen School of Computer Science and Engineering,
[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 Keras implementation and analysis pipelines of DeepATAC are available at https://github.com/hiranumn/deepatac.

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

Mingfu Shao [1,†,*], Jianzhu Ma [2,†] and Sheng Wang [3,†,*]

[1]Department of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, and
[2]School of Medicine, University of California San Diego, La Jolla, CA 92093, and
[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.

[†]The three authors contribute equally to this work.
[*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Reconstructing the full-length expressed transcripts (*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 novelly 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[1], Siddhartha Jain[2], Hannah Kim[3], Ziv Bar-Joseph[1,3,*]
[1]Machine Learning Department, [2]Computer Science Department, [3]Computational Biology Department
School of Computer Science, Carnegie Mellon University
[*] *Corresponding author, zivbj@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.