

Application of deep learning in computational biology

Hongyu & Wendy

04/19/2019

Machine Learning in CompBio

- Machine learning(supervised and unsupervised methods) have been very popular in computational biology
- Historical context
- Different methodology: From predictions to insights
- Deep Learning also becomes popular in recent years

Outline

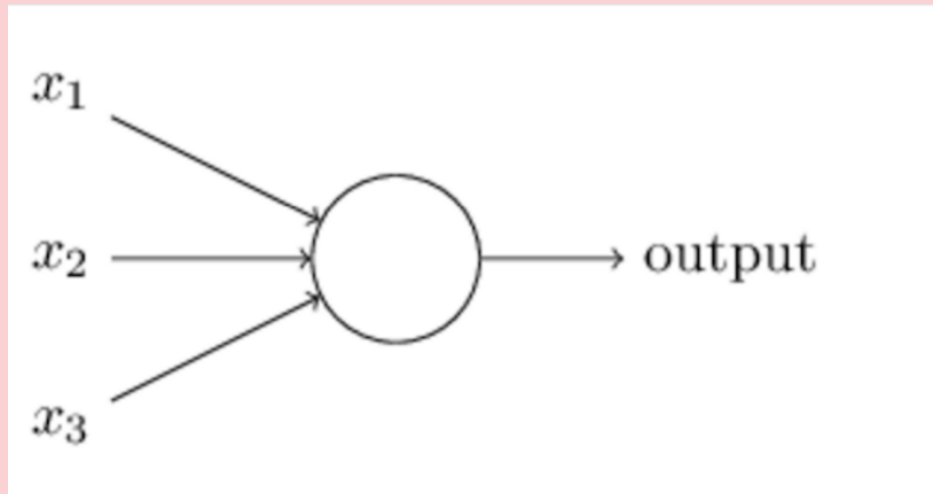
- Feedforward neural network
- Convolutional neural network (CNN)
- Recurrent neural network (RNN)
- Generative Adversarial Network (GAN)
- Deep reinforcement learning

Feedforward neural network

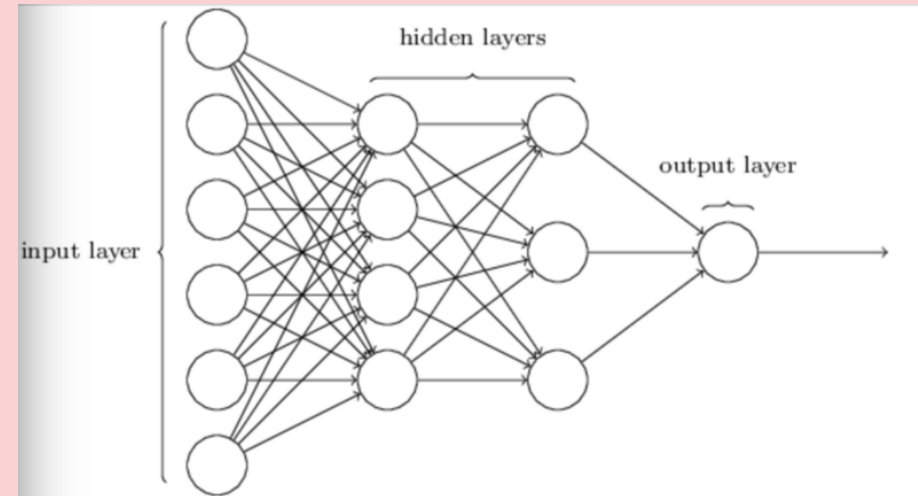
- Simplest type of neural network
- Connections don't form any cycle

One single neuron

f is a non-linear activation function



Multi-layer neural network



Application of FNN

- Genome annotation

Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome

Martin G. Reese *

Berkeley *Drosophila* Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3200, USA

Received 4 December 2000; accepted 8 May 2001

Abstract

Computational methods for automated genome annotation are critical to understanding and interpreting the bewildering mass of genomic sequence data presently being generated and released. A neural network model of the structural and compositional properties of a eukaryotic core promoter region has been developed and its application for analysis of the *Drosophila melanogaster* genome is presented. The model uses a time-delay architecture, a special case of a feed-forward neural network. The structure of this model allows for variable spacing between functional binding sites, which is known to play a key role in the transcription initiation process. Application of this model to a test set of core promoters not only gave better discrimination of potential promoter sites than previous statistical or neural network models, but also revealed indirectly subtle properties of the transcription initiation signal. When tested in the *Adh* region of 2.9 Mbases of the *Drosophila* genome, the neural network for promoter prediction (NNPP) program that incorporates the time-delay neural network model gives a recognition rate of 75% (69/92) with a false positive rate of 1/547 bases. The present work can be regarded as one of the first intensive studies that applies novel gene regulation technologies to the identification of the complex gene regulation sites in the genome of *Drosophila melanogaster*. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Neural networks; Genome annotation; Promoter recognition; DNA sequence analysis; *Drosophila melanogaster*

Li et al. BMC Bioinformatics (2018) 19:202
<https://doi.org/10.1186/s12859-018-2187-1>

BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access



Genome-wide prediction of cis-regulatory regions using supervised deep learning methods

Yifeng Li^{1,2} , Wenqiang Shi¹ and Wyeth W. Wasserman^{1*}

Abstract

Background: In the human genome, 98% of DNA sequences are non-protein-coding regions that were previously disregarded as junk DNA. In fact, non-coding regions host a variety of *cis*-regulatory regions which precisely control the expression of genes. Thus, identifying active *cis*-regulatory regions in the human genome is critical for understanding gene regulation and assessing the impact of genetic variation on phenotype. The developments of high-throughput sequencing and machine learning technologies make it possible to predict *cis*-regulatory regions genome wide.

Results: 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 based on supervised deep learning approaches 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).

Conclusion: The predicted annotations of *cis*-regulatory regions will provide broad utility for genome interpretation from functional genomics to clinical applications. The DECRES model demonstrates potentials of deep learning technologies when combined with high-throughput sequencing data, and inspires the development of other advanced neural network models for further improvement of genome annotations.

Keywords: *cis*-regulatory region, Enhancer, Promoter, Deep learning

Application of FNN

- Disease diagnosis

Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection

I.S.Isa, Z.Saad, S.Omar, M.K.Osman, K.A.Ahmad
Faculty of Electrical Engineering
Universiti Teknologi Mara (UiTM)
Kampus Pulau Pinang, 13500
Pmtg. Pauh, P.Pinang, Malaysia
izasazanita@ppinang.uitm.edu.my

H.A.Mat Sakim
School of Electrical and Electronics Engineering
Universiti sains Malaysia Kampus Transkrian
14300 Nibong Tebal, Pulau Pinang, Malaysia
amyli@eng.usm.my

Abstract –This paper presents a comparison study of various MLP activation functions for detection and classification problems. The most well-known (Artificial Neural Network) ANN architecture is the Multilayer Perceptron (MLP) network which is widely used for solving problems related to detection and data classifications. Activation function is one of the elements in MLP architecture. Selection of the activation functions in the MLP network plays an essential role on the network performance. A lot of studies have been conducted by reseachers to investigate special activation function to solve different kind of problems. Therefore, this paper intends to investigate the activation functions in MLP networks in terms of the accuracy performances. The activation functions under investigation are sigmoid, hyperbolic tangent, neuronal, logarithmic, sinusoidal and exponential. Medical diagnosis data from two case studies; thyroid disease and breast cancer, have been used to test the performance of the MLP network. The MLP networks are trained using Back Propagation learning algorithm. The performance of the MLP networks are calculated based on the percentage of correct classification. The results show that the hyperbolic tangent function in MLP network had the capability to produce the highest accuracy for detecting and hence classifying breast cancer data. Meanwhile, for thyroid disease classification, neuronal function is the most suitable function that performed the highest accuracy in MLP network.

introduce nonlinearity into the network. The selection of activation function might significantly affect the performance of a training algorithm. Some researchers have investigated to find special activation function to simplify the network structure and to accelerate convergence time [7]. An activation function for MLP network with back propagation algorithm should have several important characteristics. It should be continuous, differentiable and monotonically non-decreasing [1]. Table 1 summarizes various activation functions used in neural network.

Research on effect of activation function for neural network has received a lot of attention in the past literatures. In [14], a Cosine-Modulated Gaussian activation function for Hyper-Hill neural networks has been proposed. The study compared the Cosine-Modulated Gaussian, hyperbolic tangent, sigmoid and symsgmoid function in cascade correlation network to solve sonar benchmark problem. Joarder and Aziz [6] proved that logarithmic function is able to accelerate back propagation learning or network convergence. The study has solved XOR problem, character recognition, machine learning database and encoder problem using MLP network with back propagation learning. Wong *et al* [11] investigated the neuronal function for network convergence and pruning performance. Periodic and

Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks

Javed Khan^{1,2}, Jun S. Wei¹, Markus Ringnér^{1,3}, Lao H. Saal¹, Marc Ladanyi⁴, Frank Westermann⁵, Frank Berthold⁶, Manfred Schwab⁵, Cristina R. Antonescu⁴, Carsten Peterson³, and Paul S. Meltzer¹

1Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA

2Pediatric Oncology Branch, Advanced Technology Center, National Cancer Institute, Gaithersburg, Maryland, USA

3Complex Systems Division, Department of Theoretical Physics, Lund University, Lund, Sweden

4Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA

5Department of Cytogenetics, German Cancer Research Center, Heidelberg, Germany

6Department of Pediatrics, Klinik für Kinderheilkunde der Universität zu Köln, Köln, Germany

Abstract

The purpose of this study was to develop a method of classifying cancers to specific diagnostic categories based on their gene expression signatures using artificial neural networks (ANNs). We trained the ANNs using the small, round blue-cell tumors (SRBCTs) as a model. These cancers belong to four distinct diagnostic categories and often present diagnostic dilemmas in clinical practice. The ANNs correctly classified all samples and identified the genes most relevant to the classification. Expression of several of these genes has been reported in SRBCTs, but most have not been associated with these cancers. To test the ability of the trained ANN models to recognize SRBCTs, we analyzed additional blinded samples that were not previously used for the training procedure, and correctly classified them in all cases. This study demonstrates the potential applications of these methods for tumor diagnosis and the identification of candidate targets for therapy.

Application of FNN

- Pattern recognition

Clustered-Hybrid Multilayer Perceptron network for pattern recognition application

Nor Ashidi Mat Isa*, Wan Mohd Fahmi Wan Mamat¹

Imaging and Intelligent System Research Team (ISRT), School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia

ARTICLE INFO

Article history:

Received 28 February 2008

Received in revised form 27 April 2010

Accepted 28 April 2010

Available online 5 May 2010

Keywords:

Clustered-Hybrid Multilayer Perceptron network

Clustered-Modified Recursive Prediction Error

Pattern Recognition

Radial Basis Function

Clustering Algorithm

Neural network

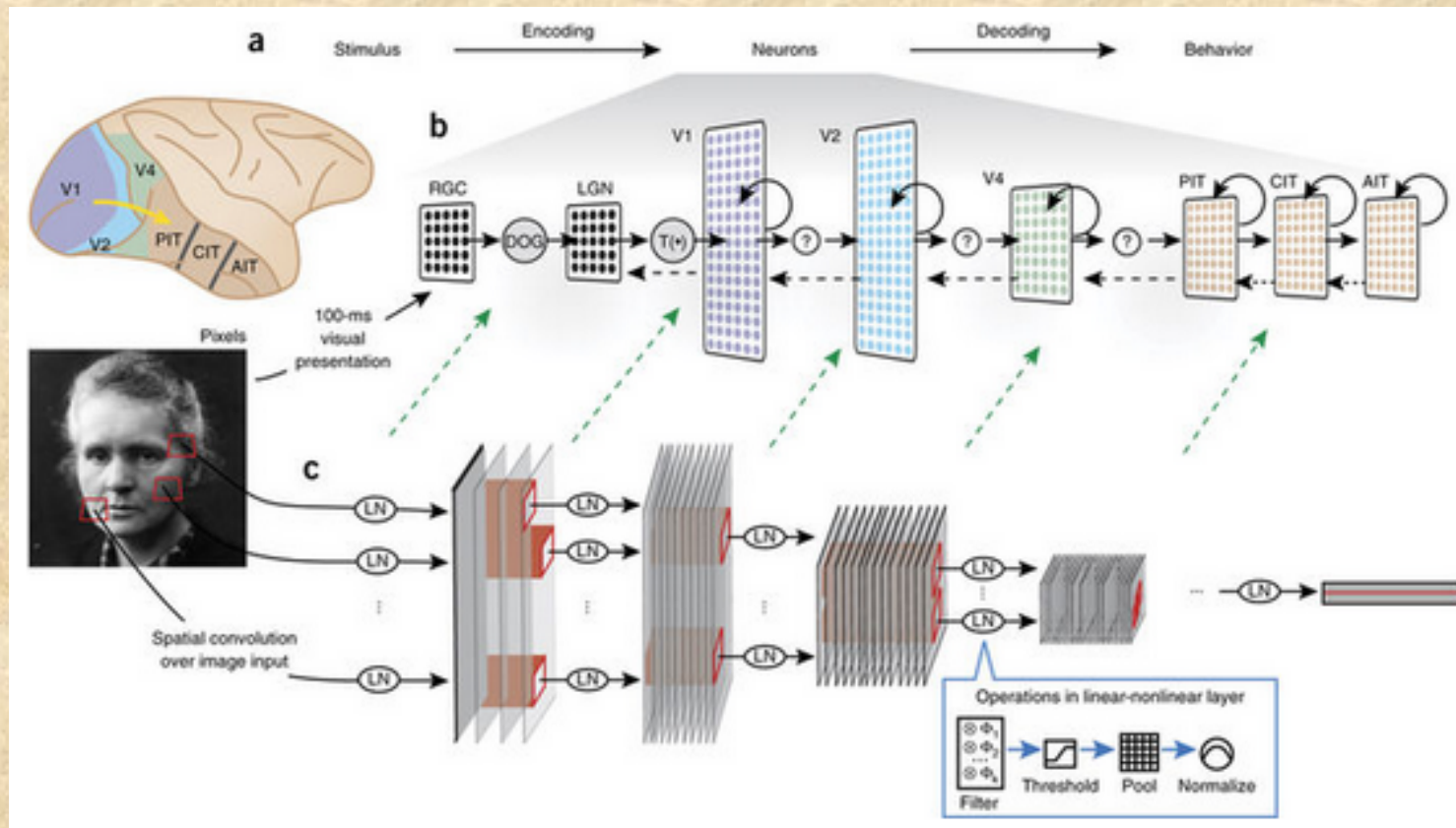
ABSTRACT

This paper introduces a modified version of the Hybrid Multilayer Perceptron (HMLP) network to improve the performance of the conventional HMLP network. We adopted the Clustering Algorithm from the Radial Basis Function (RBF) network architecture and incorporated it into the conventional HMLP network architecture. The modified model is called Clustered-Hybrid Multilayer Perceptron (Clustered-HMLP) network. The proposed Clustered-HMLP network architecture is trained using modified training algorithm called Clustered-Modified Recursive Prediction Error (Clustered-MRPE). The capability of the Clustered-HMLP network with Clustered-MRPE training algorithm is demonstrated using seven benchmark datasets from the University of California at Irvine (UCI) machine learning repository (i.e. Iris, Ionosphere, Pima Indian Diabetes, Wine, Lung Cancer, Hayes-Roth and Glass) and compared with the performance of other twelve classifiers reported in literature. Further, the new network is implemented to model a Transformer Fault Diagnosis System and Aggregate Shape Identification System. The results indicate that the proposed Clustered-HMLP network outperforms other eleven classifiers and provides a significant improvement to the conventional HMLP network for pattern recognition application.

© 2010 Elsevier B.V. All rights reserved.

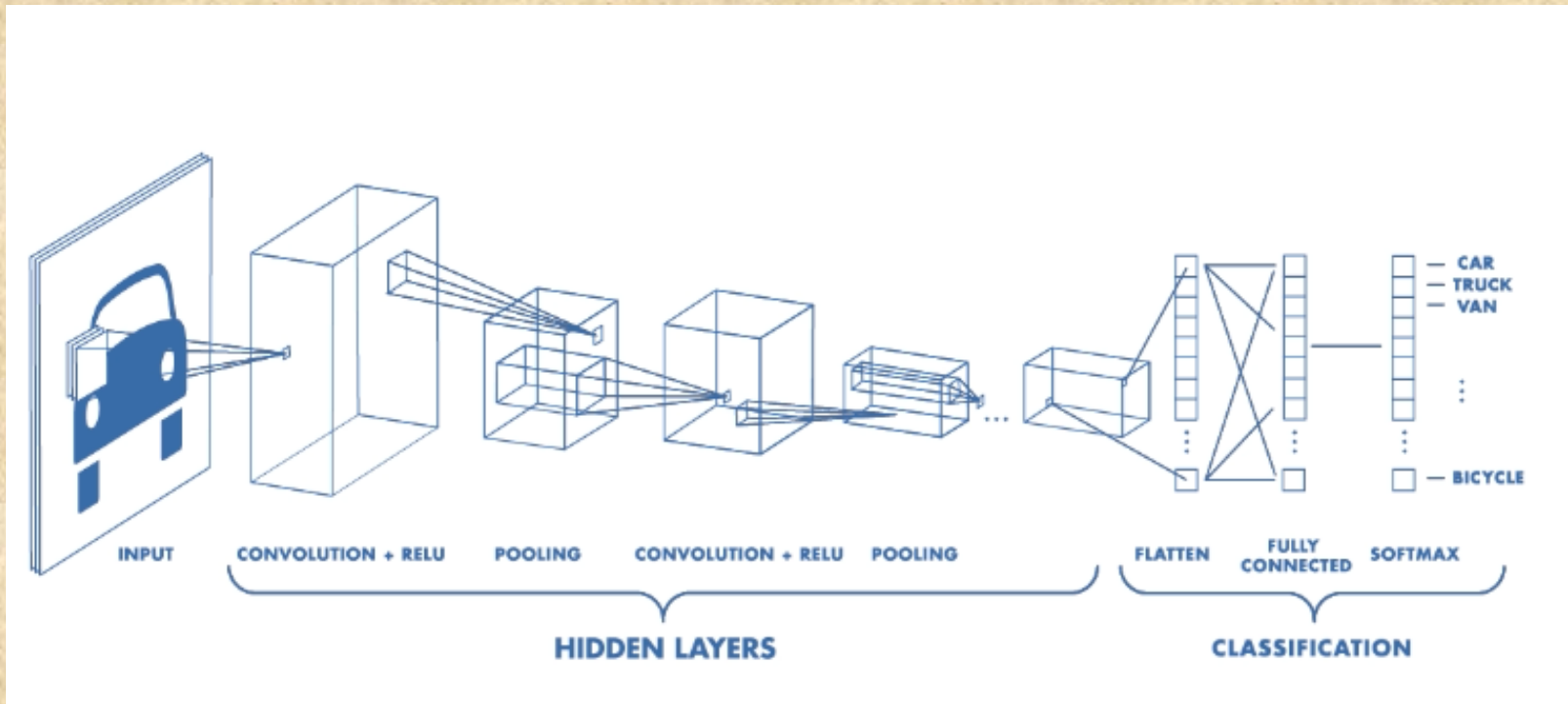
Convolutional Neural Network

- Inspired by visual cortex
- Usually used in image analysis



Convolutional Neural Network

- **Convolutional layer:** feature extraction
- **Pooling layer:** reduce dimension of data



Application of CNN

- Gene prediction

[Interdisciplinary Sciences: Computational Life Sciences](#)

pp 1–8 | [Cite as](#)

CNN-MGP: Convolutional Neural Networks for Metagenomics Gene Prediction

Authors

[Authors and affiliations](#)

Amani Al-Ajlan  , Achraf El Allali

[Open Access](#) | Original Research Article

First Online: 27 December 2018

675

Downloads

Abstract

Accurate gene prediction in metagenomics fragments is a computationally challenging task due to the short-read length, incomplete, and fragmented nature of the data. Most gene-prediction programs are based on extracting a large number of features and then applying statistical approaches or supervised classification approaches to predict genes. In our study, we introduce a convolutional neural network for metagenomics gene prediction (CNN-MGP) program that predicts genes in metagenomics fragments directly from raw DNA sequences, without the need for manual feature extraction and feature selection stages. CNN-MGP is able to learn the characteristics of coding and non-coding regions and distinguish coding and non-coding open reading frames (ORFs). We train 10 CNN models on 10 mutually exclusive datasets based on pre-defined GC content ranges. We extract ORFs from each fragment; then, the ORFs are

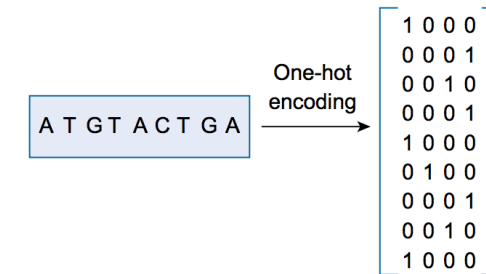


Fig. 1 One-hot Encoding for DNA sequence. Each nucleotide is represented as a one-hot vector: A = 1000, T = 0001, C = 0100, and G = 0010

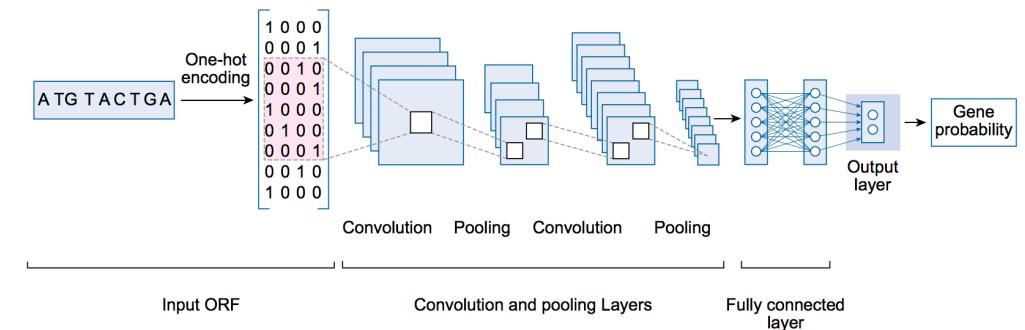



Fig. 2 CNN-MGP Architecture. First, an ORF is encoded numerically using one-hot encoding; then, a matrix of numbers is inputted into an appropriate CNN-MGP model based on its fragment GC content. The CNN-MGP model consists of six layers. The first layer is a convolutional layer with 64 filters and a filter window size of 21. The second layer is a max-pooling layer with a pool size of 2. The third

layer is a convolutional layer with 200 filters and a filter window size of 21, and the fourth layer is a max-pooling layer with a pool size of 2. Then, the output is flattened to a 1D vector before being inputted into a fully connected layer with 128 neurons. Then, the output layer produces a final gene probability

Application of CNN

- Annotation of cellular images

Convolutional neural networks for automated annotation of cellular cryo-electron tomograms

Muyuan Chen^{1,2}, Wei Dai^{2,4}, Stella Y Sun²,
Darius Jonasch², Cynthia Y He³, Michael F Schmid²,
Wah Chiu² & Steven J Ludtke² 

Cellular electron cryotomography offers researchers the ability to observe macromolecules frozen in action *in situ*, but a primary challenge with this technique is identifying molecular components within the crowded cellular environment. We introduce a method that uses neural networks to dramatically reduce the time and human effort required for subcellular annotation and feature extraction. Subsequent subtomogram classification and averaging yield *in situ* structures of molecular components of interest. The method is available in the EMAN2.2 software package.

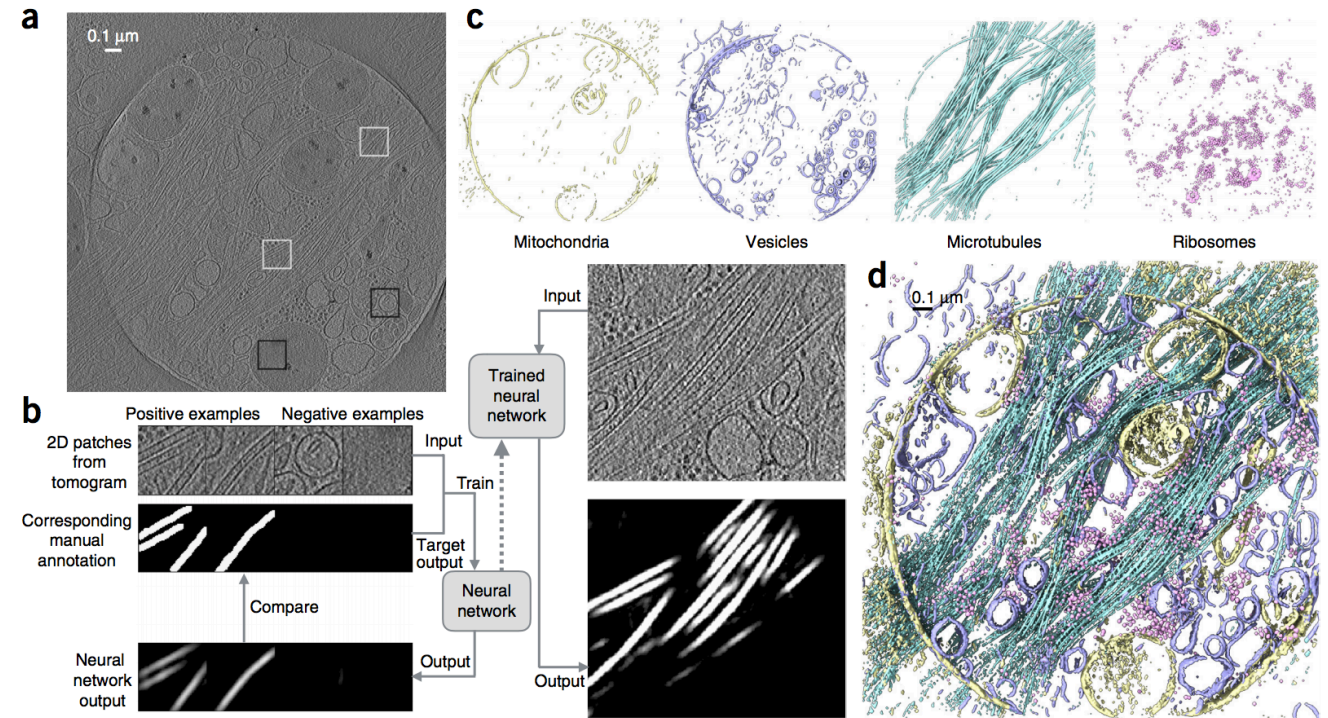


Figure 1 | Workflow of automated tomogram annotation using a PC12 cell as an example. (a) Slice view of a raw tomogram input. Locations of representative positive (white) and negative (black) examples are shown as boxes. (b) Training and annotation of a single feature. Representative 2D patches of both positive (10 total) and negative examples (86 total) are extracted and manually segmented. The time required for manual selection and annotation was ~5 min. Tomogram patches and their corresponding annotations are used to train the neural network, so that the network outputs match manual annotations. The trained neural network is applied to whole 2D slices, and the feature of interest is segmented. (c) Annotation of multiple features in a tomogram. Four neural networks are trained independently to recognize double membrane (yellow), single membrane (blue), microtubule (cyan) and ribosome (pink). (d) Masked-out density of the merged final annotation of the four features.

Application of CNN

- Tumor classification

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva^{1*}, Brett Kuprel^{1*}, Roberto A. Novoa^{2,3}, Justin Ko², Susan M. Swetter^{2,4}, Helen M. Blau⁵ & Sebastian Thrun⁶

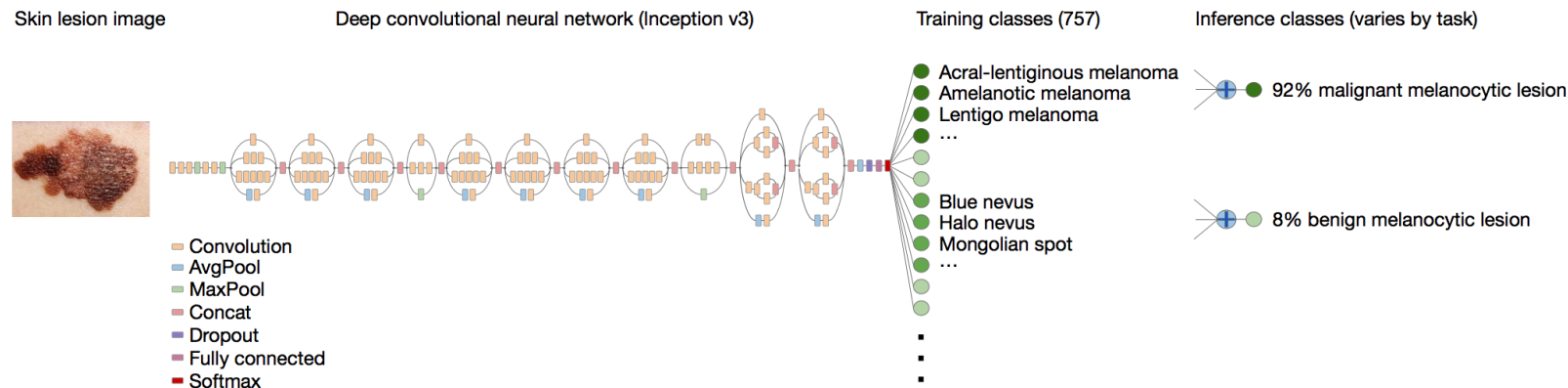
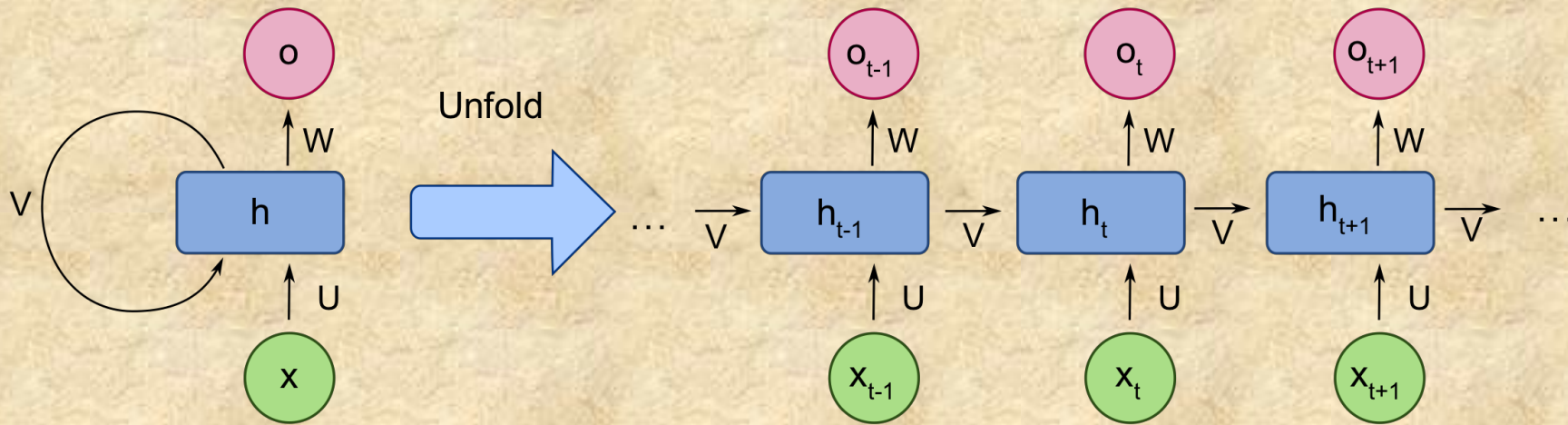


Figure 1 | Deep CNN layout. Our classification technique is a deep CNN. Data flow is from left to right: an image of a skin lesion (for example, melanoma) is sequentially warped into a probability distribution over clinical classes of skin disease using Google Inception v3 CNN architecture pretrained on the ImageNet dataset (1.28 million images over 1,000 generic object classes) and fine-tuned on our own dataset of 129,450 skin lesions comprising 2,032 different diseases. The 757 training classes are defined using a novel taxonomy of skin disease and a partitioning algorithm that maps diseases into training classes

(for example, acrolentiginous melanoma, amelanotic melanoma, lentigo melanoma). Inference classes are more general and are composed of one or more training classes (for example, malignant melanocytic lesions—the class of melanomas). The probability of an inference class is calculated by summing the probabilities of the training classes according to taxonomy structure (see Methods). Inception v3 CNN architecture reprinted from <https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html>

Recurrent neural network

- RNN has 'memory' compared with FNN
- RNN is able to exhibit temporal dynamic behavior
- Widely used in machine translation and speech recognition



Application of RNN

- Binding site prediction

RNN allows us to incorporate context information into model.

Recurrent Neural Network for Predicting Transcription Factor Binding Sites

Zhen Shen, Wenzheng Bao  & De-Shuang Huang

It is well known that DNA sequence contains a certain amount of transcription factors (TF) binding sites, and only part of them are identified through biological experiments. However, these experiments are expensive and time-consuming. To overcome these problems, some computational methods, based on k-mer features or convolutional neural networks, have been proposed to identify TF binding sites from DNA sequences. Although these methods have good performance, the context information that relates to TF binding sites is still lacking. Research indicates that standard recurrent neural networks (RNN) and its variants have better performance in time-series data compared with other models. In this study, we propose a model, named KEGRU, to identify TF binding sites by combining Bidirectional Gated Recurrent Unit (GRU) network with k-mer embedding. Firstly, DNA sequences are divided into k-mer sequences with a specified length and stride window. And then, we treat each k-mer as a word and pre-trained word representation model through word2vec algorithm. Thirdly, we construct a deep bidirectional GRU model for feature learning and classification. Experimental results have shown that our method has better performance compared with some state-of-the-art methods. Additional experiments about embedding strategy show that k-mer embedding will be helpful to enhance model performance. The robustness of KEGRU is proved by experiments with different k-mer length, stride window and embedding vector dimension.

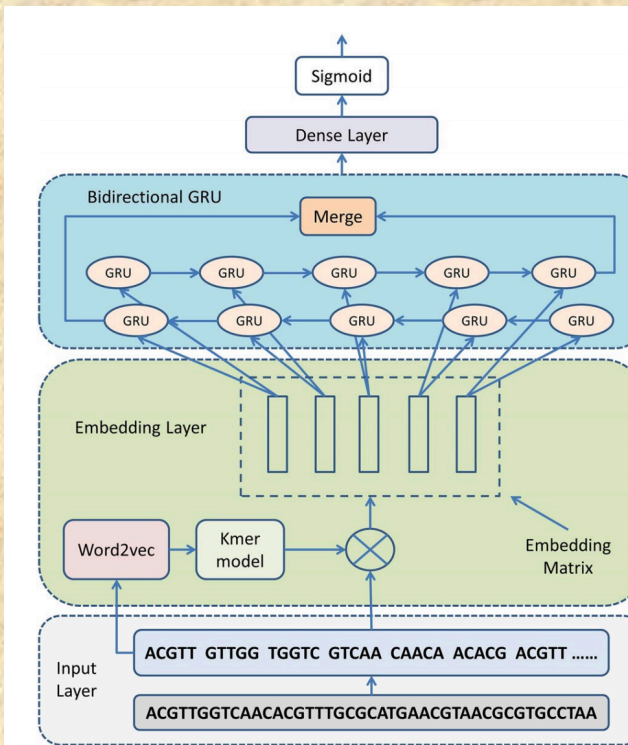


Figure 5. The basic architectural structure of our model KEGRU. (1) We first built the k-mer corpus, which consists of a number of k-mer sequence built by splitting DNA sequence. (2) Based on the k-mer corpus built at first step, we use the pre-trained model word2vec to learn the k-mer embedding vectors. All k-mer vectors are stacked into the embedding matrix that will be used to initialize the embedding layer. (3) We use bidirectional GRU network to solve long-range dependencies problem and to learn feature information from input k-mer sequence. (4) The prediction results were generated by the dense layer and the sigmoid layer, and then we use a loss function to compare the prediction results with the true target labels.

Application of RNN

• Identification of miRNA

Deep Recurrent Neural Network-Based Identification of Precursor microRNAs

Seunghyun Park

Electrical and Computer Engineering
Seoul National University
Seoul 08826, Korea
School of Electrical Engineering
Korea University
Seoul 02841, Korea

Seonwoo Min

Electrical and Computer Engineering
Seoul National University
Seoul 08826, Korea

Hyun-Soo Choi

Electrical and Computer Engineering
Seoul National University
Seoul 08826, Korea

Sungroh Yoon*

Electrical and Computer Engineering
Seoul National University
Seoul 08826, Korea
sryoon@snu.ac.kr

Abstract

MicroRNAs (miRNAs) are small non-coding ribonucleic acids (RNAs) which play key roles in post-transcriptional gene regulation. Direct identification of mature miRNAs is infeasible due to their short lengths, and researchers instead aim at identifying precursor miRNAs (pre-miRNAs). Many of the known pre-miRNAs have distinctive stem-loop secondary structure, and structure-based filtering is usually the first step to predict the possibility of a given sequence being a pre-miRNA. To identify new pre-miRNAs that often have non-canonical structure, however, we need to consider additional features other than structure. To obtain such additional characteristics, existing computational methods rely on manual feature extraction, which inevitably limits the efficiency, robustness, and generalization of computational identification. To address the limitations of existing approaches, we propose a pre-miRNA identification method that incorporates (1) a deep recurrent neural network (RNN) for automated feature learning and classification, (2) multimodal architecture for seamless integration of prior knowledge (secondary structure), (3) an attention mechanism for improving long-term dependence modeling, and (4) an RNN-based class activation mapping for highlighting the learned representations that can contrast pre-miRNAs and non-pre-miRNAs. In our experiments with recent benchmarks, the proposed approach outperformed the compared state-of-the-art alternatives in terms of various performance metrics.

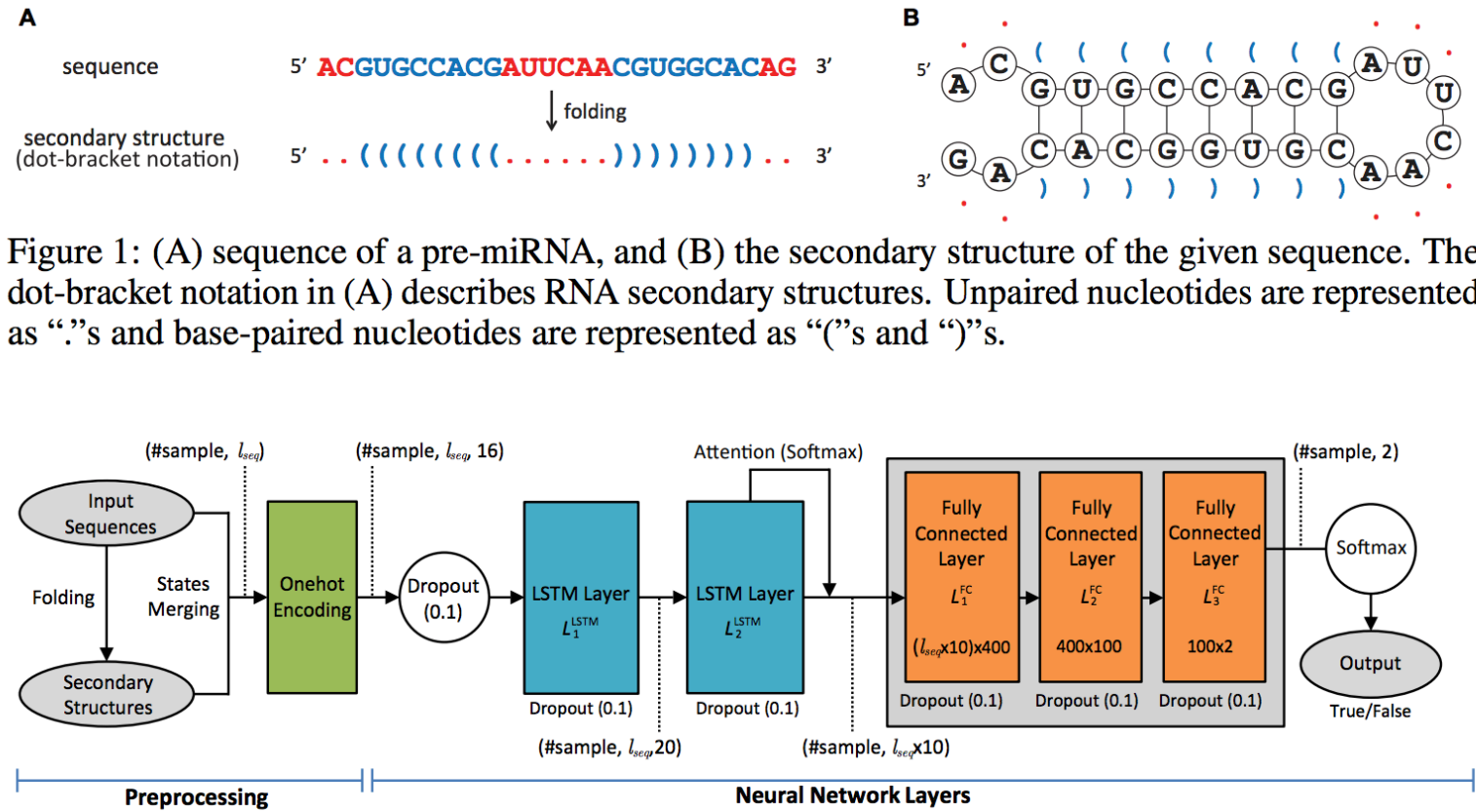


Figure 2: Overview of our method: #sample is the number of input sequences and l_{seq} is the maximum length of the input sequence. The dimension of intermediate data is labeled (#sample, l_{seq} , 16).

Application of RNN

- Gene regulatory network reconstruction

Recurrent neural network based hybrid model for reconstructing gene regulatory network

Khalid Raza*, Mansaf Alam

Department of Computer Science, Jamia Millia Islamia (Central University), New Delhi-110025, India



ARTICLE INFO

Article history:

Received 8 January 2016
Received in revised form 1 May 2016
Accepted 13 August 2016
Available online 16 August 2016

Keywords:

Recurrent neural network
Gene regulatory network model
Gene expression
Kalman filter

ABSTRACT

One of the exciting problems in systems biology research is to decipher how genome controls the development of complex biological system. The gene regulatory networks (GRNs) help in the identification of regulatory interactions between genes and offer fruitful information related to functional role of individual gene in a cellular system. Discovering GRNs lead to a wide range of applications, including identification of disease related pathways providing novel tentative drug targets, helps to predict disease response, and also assists in diagnosing various diseases including cancer. Reconstruction of GRNs from available biological data is still an open problem. This paper proposes a recurrent neural network (RNN) based model of GRN, hybridized with generalized extended Kalman filter for weight update in backpropagation through time training algorithm. The RNN is a complex neural network that gives a better settlement between biological closeness and mathematical flexibility to model GRN; and is also able to capture complex, non-linear and dynamic relationships among variables. Gene expression data are inherently noisy and Kalman filter performs well for estimation problem even in noisy data. Hence, we applied non-linear version of Kalman filter, known as generalized extended Kalman filter, for weight update during RNN training. The developed model has been tested on four benchmark networks such as DNA SOS repair network, IRMA network, and two synthetic networks from DREAM Challenge. We performed a comparison of our results with other state-of-the-art techniques which shows superiority of our proposed model. Further, 5% Gaussian noise has been induced in the dataset and result of the proposed model shows negligible effect of noise on results, demonstrating the noise tolerance capability of the model.

© 2016 Elsevier Ltd. All rights reserved.

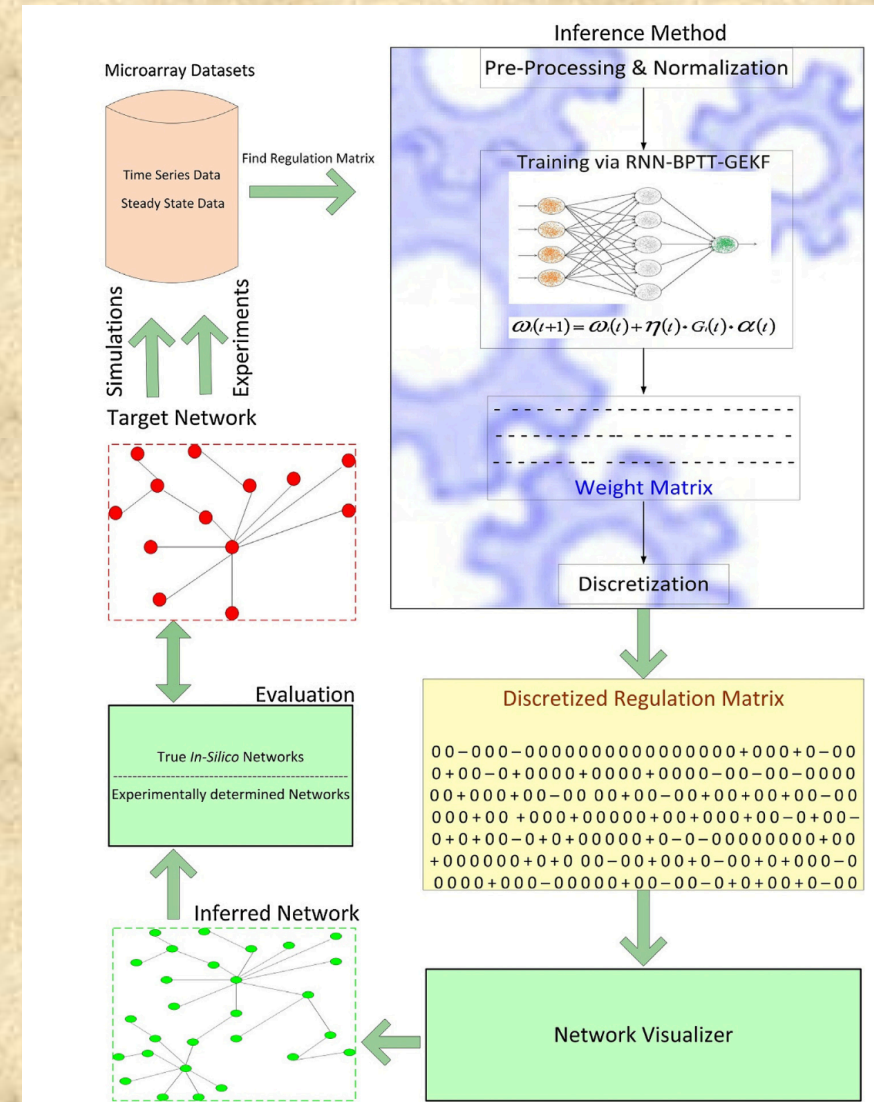
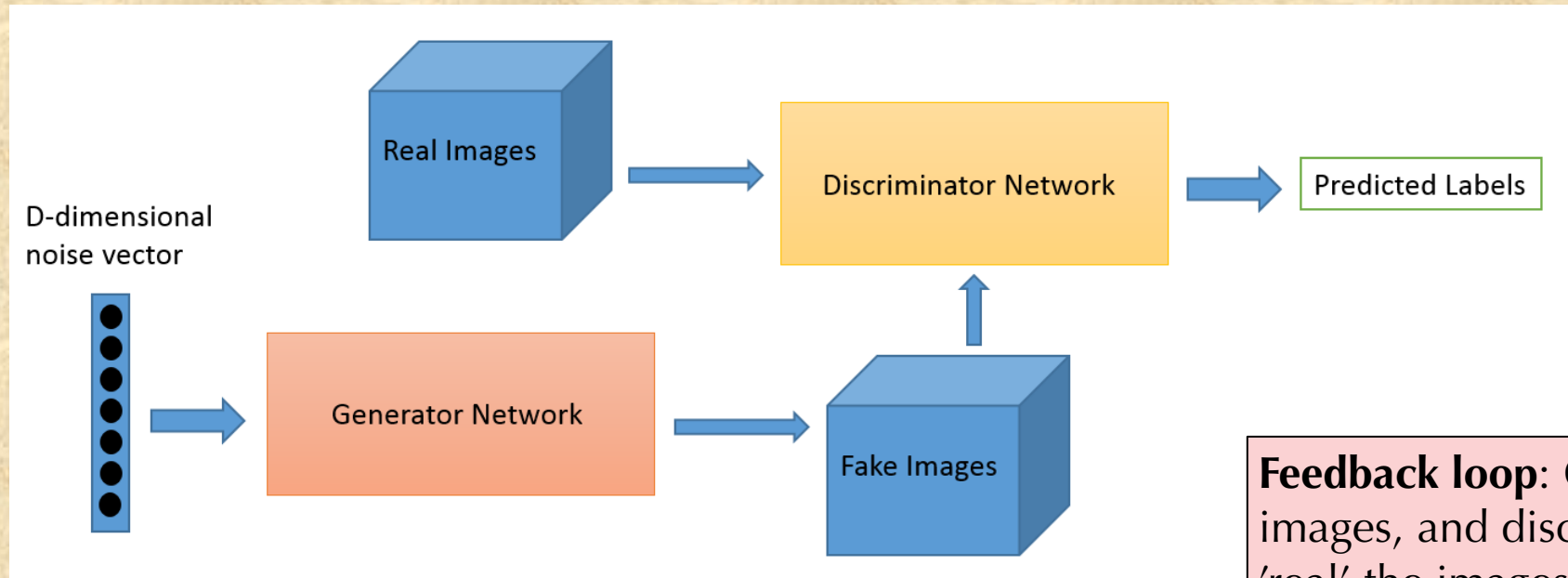


Fig. 2. Proposed model architecture.

Generative Adversarial Network (GAN)

- All models above are discriminative models.
- GAN is able to generate new samples that look authentic.

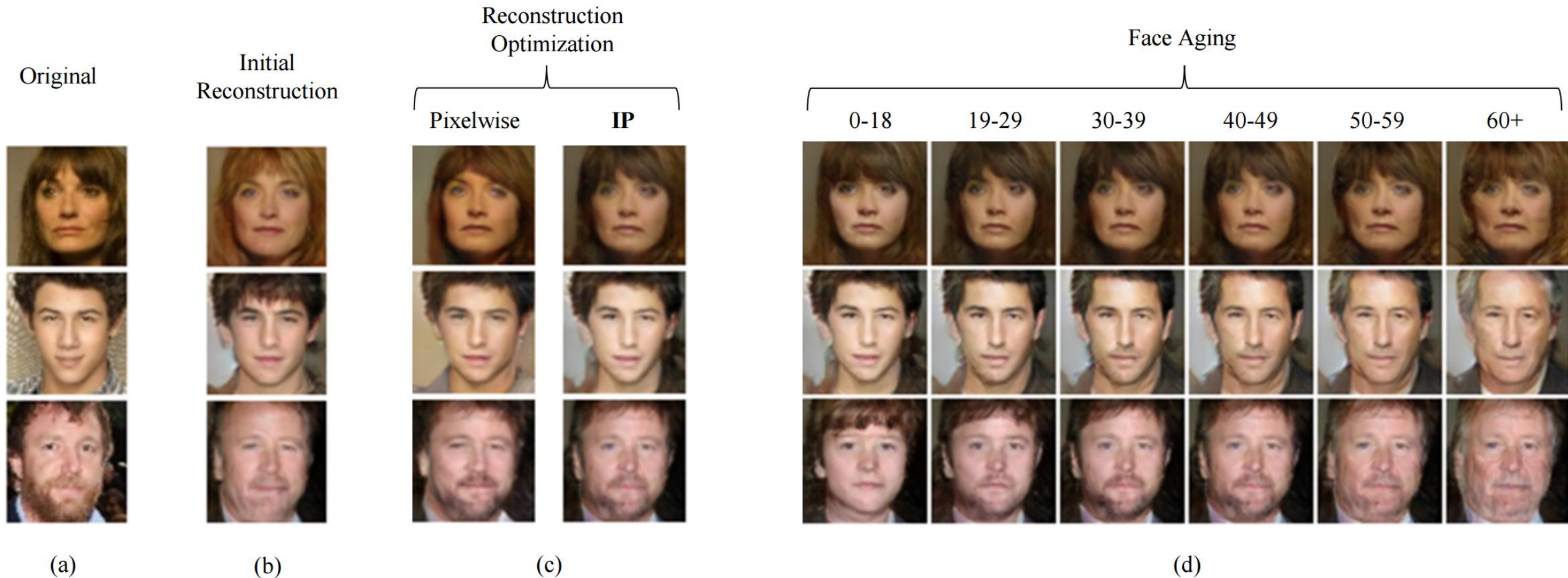


Feedback loop: Generator generates new images, and discriminator tells it how 'real' the images are.

Just like the game of cat and mouse!

Generative Adversarial Network (GAN)

- Some fun applications



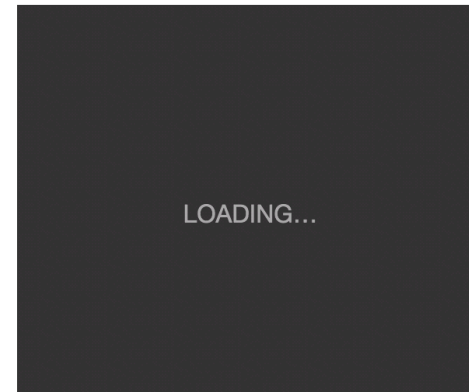
Generative Adversarial Network (GAN)

- Some fun applications

FAKE video: <https://www.youtube.com/watch?v=UXjWWy6iTVo>

Carnegie Mellon researchers create the most convincing deepfakes yet

KYLE WIGGERS @KYLE_L_WIGGERS AUGUST 16, 2018 8:12 AM



MOST READ



Application of GAN

- More patient data

Generating Multi-label Discrete Patient Records using Generative Adversarial Networks

Edward Choi¹

MP2893@GATECH.EDU

Siddharth Biswal¹

SBISWAL7@GATECH.EDU

Bradley Malin²

BRADLEY.MALIN@VANDERBILT.EDU

Jon Duke¹

JON.DUKE@GATECH.EDU

Walter F. Stewart³

STEWARWF@SUTTERHEALTH.ORG

Jimeng Sun¹

JSUN@CC.GATECH.EDU

¹GEORGIA INSTITUTE OF TECHNOLOGY ² VANDERBILT UNIVERSITY ³ SUTTER HEALTH

Abstract

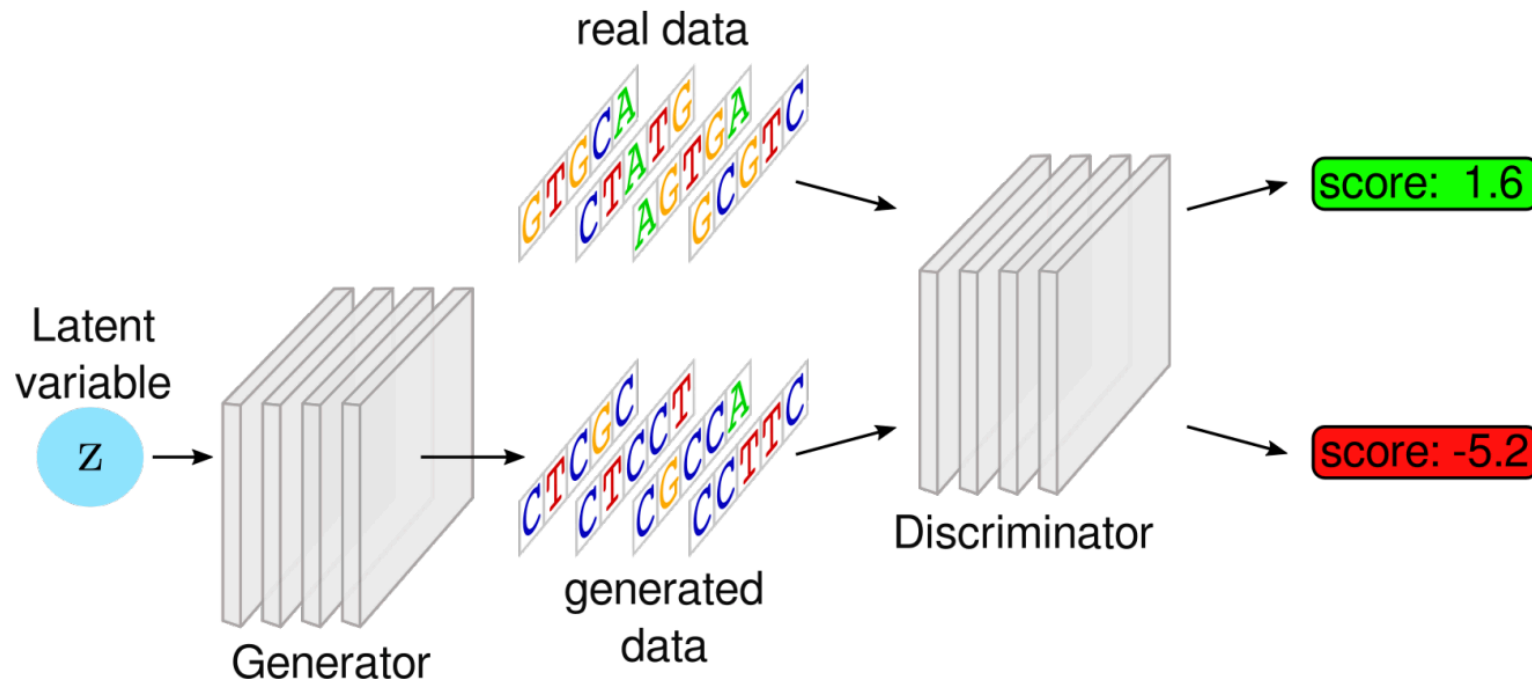
Access to electronic health record (EHR) data has motivated computational advances in medical research. However, various concerns, particularly over privacy, can limit access to and collaborative use of EHR data. Sharing synthetic EHR data could mitigate risk.

In this paper, we propose a new approach, medical Generative Adversarial Network (medGAN), to generate realistic synthetic patient records. Based on input real patient records, medGAN can generate high-dimensional discrete variables (e.g., binary and count features) via a combination of an autoencoder and generative adversarial networks. We also propose minibatch averaging to efficiently avoid mode collapse, and increase the learning efficiency with batch normalization and shortcut connections. To demonstrate feasibility, we showed that medGAN generates synthetic patient records that achieve comparable performance to real data on many experiments including distribution statistics, predictive modeling tasks and a medical expert review. We also empirically observe a limited privacy risk in both identity and attribute disclosure using medGAN.

Application of GAN

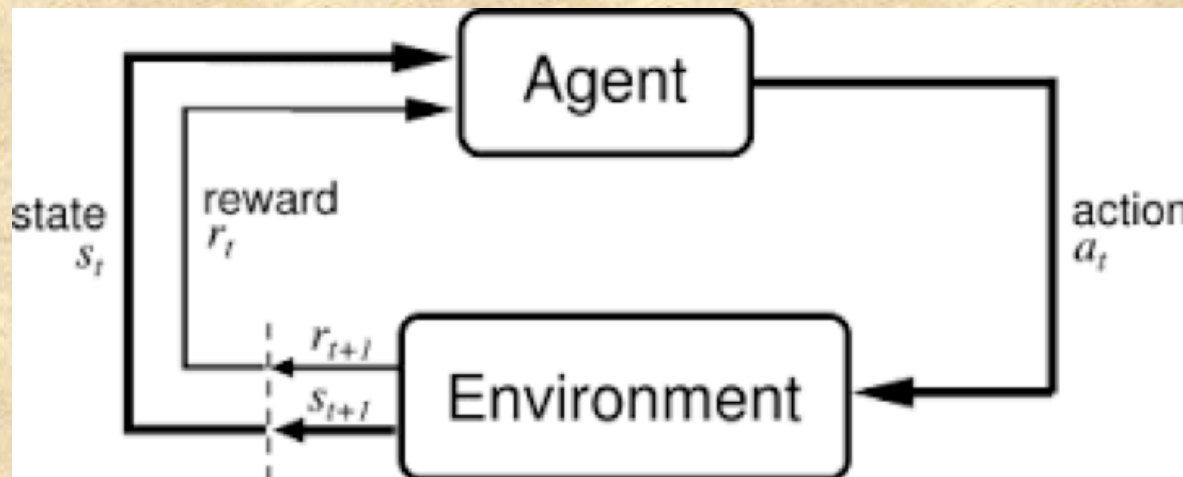
- More DNA!

Generating and designing DNA with deep generative models



Deep Reinforcement Learning

- Combines deep learning and reinforcement learning
- Reinforcement learning is very similar to animal learning: An agent learns by interacting with the environment, picking actions that gives higher reward.
- Deep learning allows RL to generalize better for unseen states.



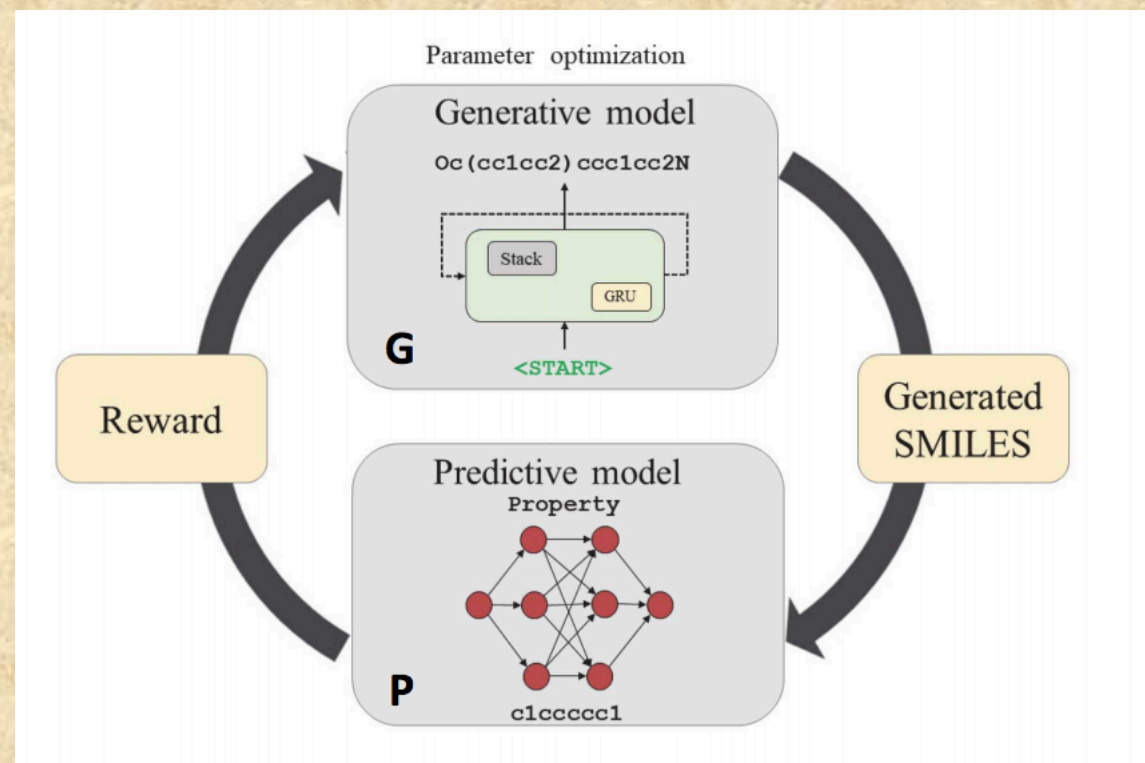
Application of deep RL

- De novo drug discovery

Deep reinforcement learning for de novo drug design

Mariya Popova^{1,2,3}, Olexandr Isayev^{1*}, Alexander Tropsha^{1*}

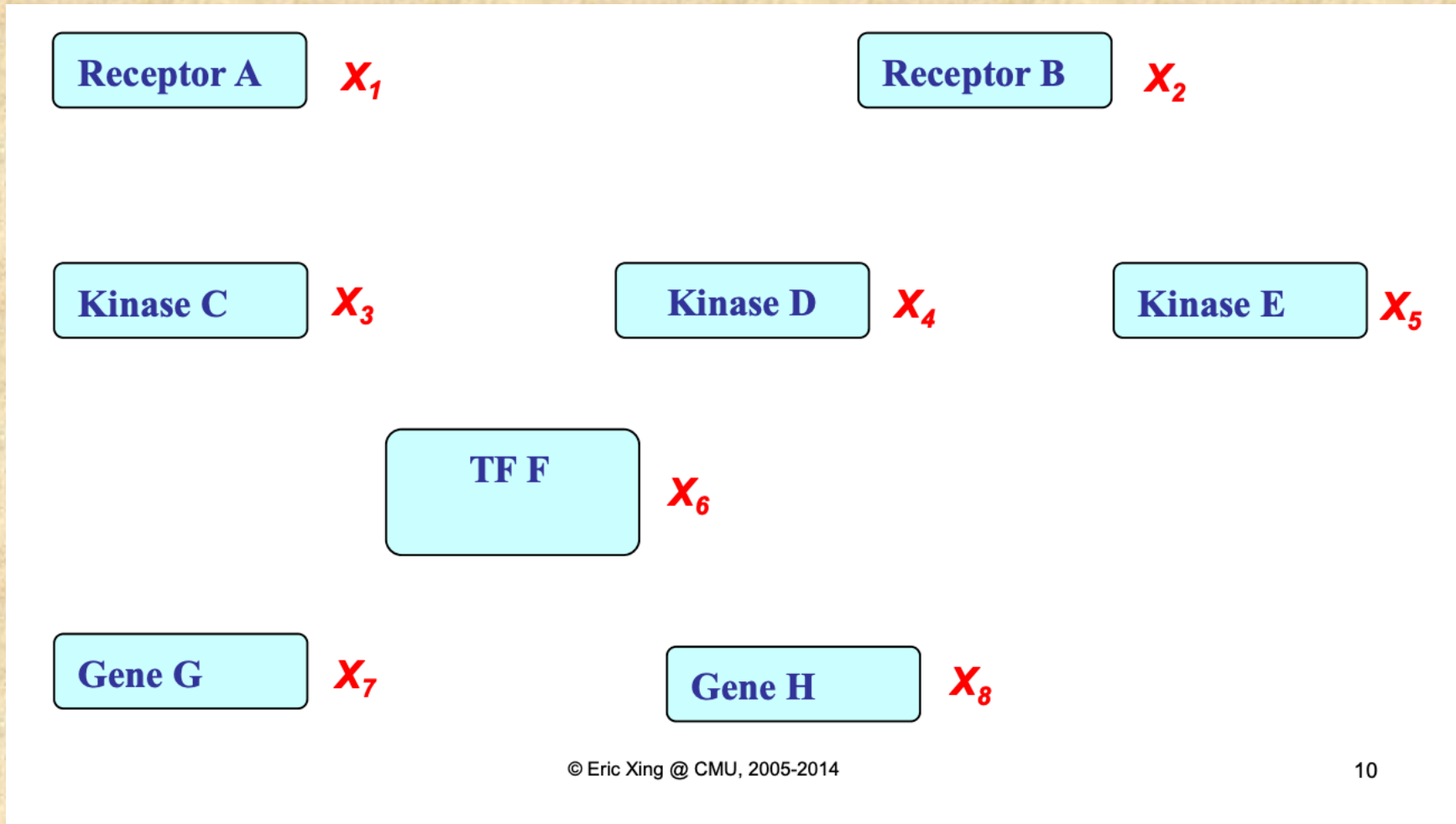
We have devised and implemented a novel computational strategy for de novo design of molecules with desired properties termed ReLeaSE (Reinforcement Learning for Structural Evolution). On the basis of deep and reinforcement learning (RL) approaches, ReLeaSE integrates two deep neural networks—generative and predictive—that are trained separately but are used jointly to generate novel targeted chemical libraries. ReLeaSE uses simple representation of molecules by their simplified molecular-input line-entry system (SMILES) strings only. Generative models are trained with a stack-augmented memory network to produce chemically feasible SMILES strings, and predictive models are derived to forecast the desired properties of the de novo-generated compounds. In the first phase of the method, generative and predictive models are trained separately with a supervised learning algorithm. In the second phase, both models are trained jointly with the RL approach to bias the generation of new chemical structures toward those with the desired physical and/or biological properties. In the proof-of-concept study, we have used the ReLeaSE method to design chemical libraries with a bias toward structural complexity or toward compounds with maximal, minimal, or specific range of physical properties, such as melting point or hydrophobicity, or toward compounds with inhibitory activity against Janus protein kinase 2. The approach proposed herein can find a general use for generating targeted chemical libraries of novel compounds optimized for either a single desired property or multiple properties.



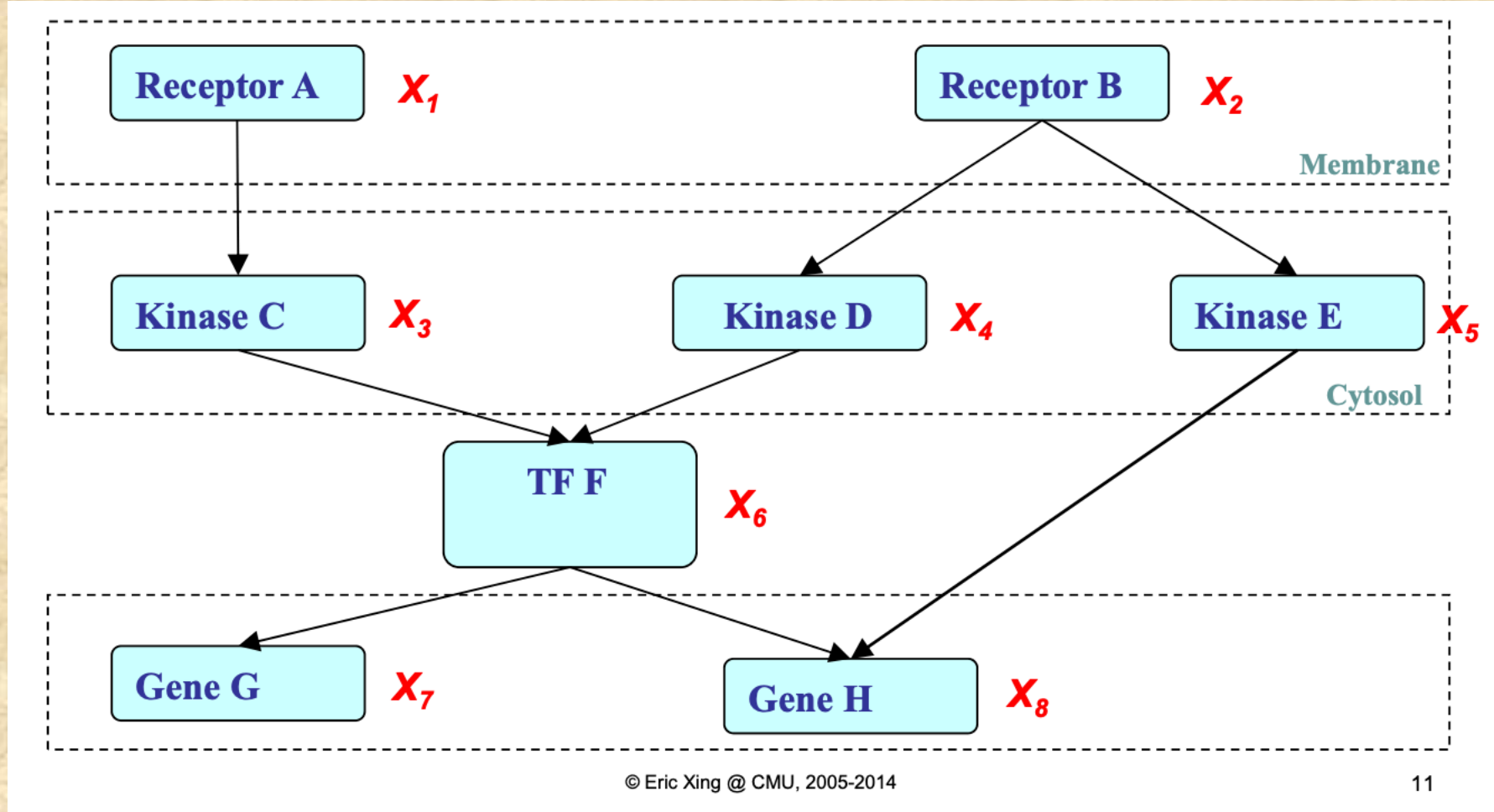
Some Extra Material: Graphical Models

- A (wildly) popular approach for ML in computational biology
- A lot of early(and current!) investigators in this field are also computational biologists

Graphical Models(Slides from 10-708 14S)



Graphical Models(Slides from 10-708 14S)



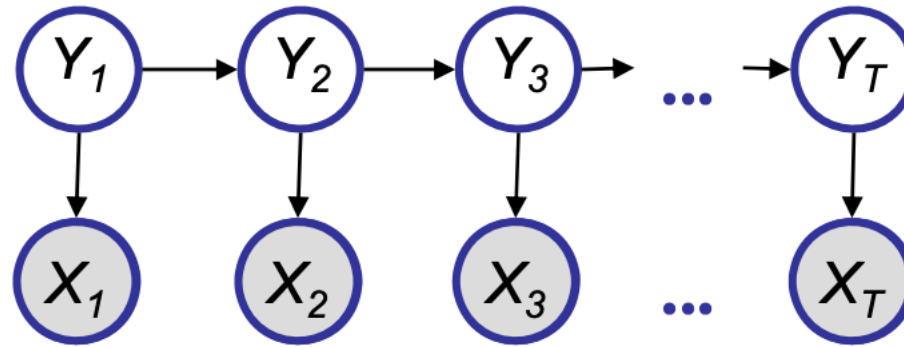
Graphical Models(Slides from 10-708 14S)

The underlying source:

Speech signal
genome function
dice

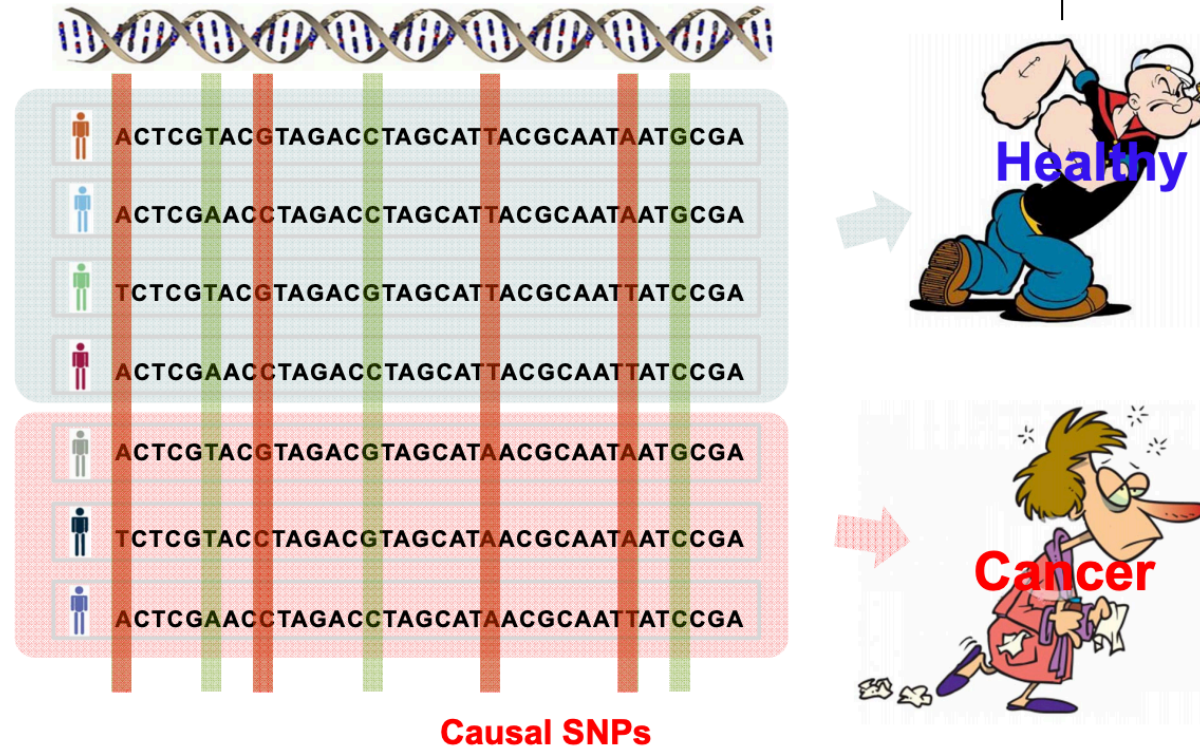
The sequence:

Phonemes
DNA sequence
sequence of rolls



Graphical Models(Slides from 10-708 14S)

Genetic Basis of Complex Diseases



Graphical Models(Slides from 10-708 14S)

Genetic Basis of Complex Diseases

