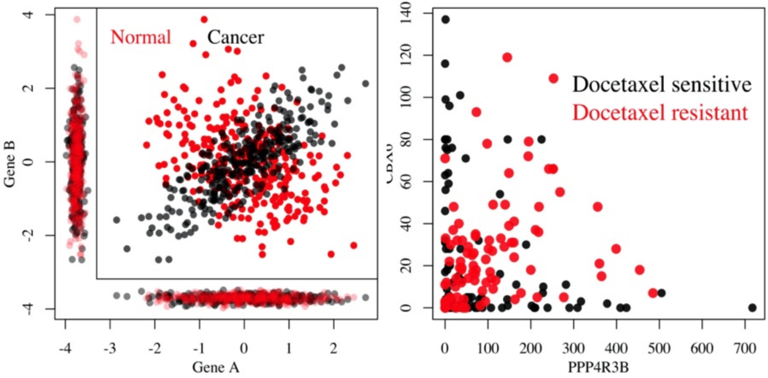
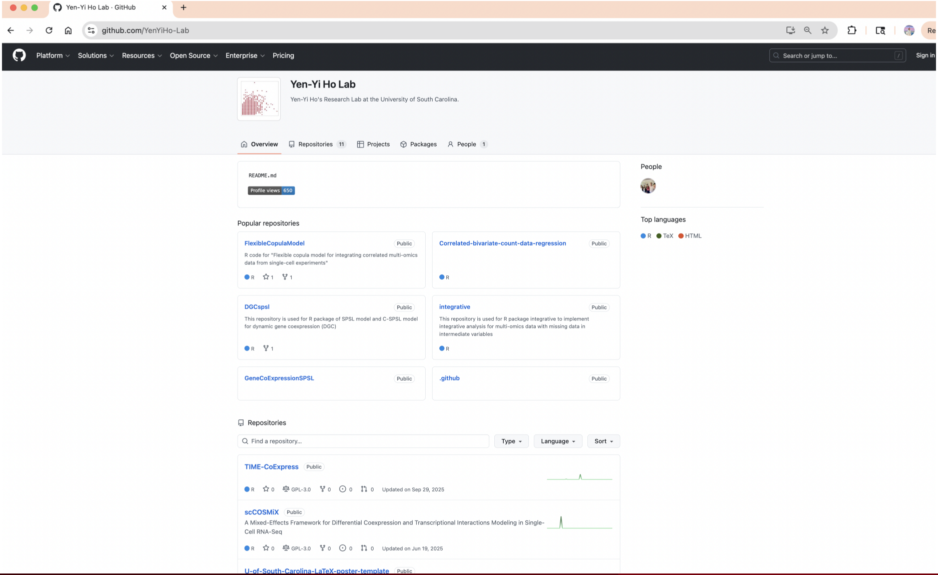
Lecture 12: Gene Co-Expression Analysis

Yen-Yi Ho

Department of Statistics

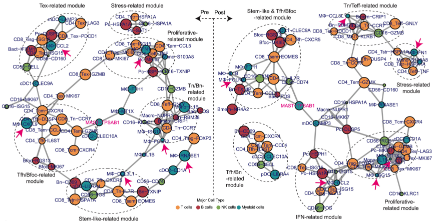

Normal
Cancer
Gene B
Gene A
4
2
0
-2
-4
-3
-1
1
Docetaxel sensitive
Docetaxel resistant
PPP4R3B
0
20
40
60
80
100
120
140
200
300
400
500
600
700

Co-Expression Analysis



scCOSMIX & Time-CoExpress

- Single-cell RNA-seq enables gene–gene interaction analysis at cellular resolution
- Many analyses focus on **mean expression**, ignoring interactions
- Gene co-expression can:
  - vary across conditions (treatment, cell type)
  - evolve continuously along pseudotime
- Key challenges:
  - zero inflation and over-dispersion
  - hierarchical (multi-patient) designs
  - non-linear, covariate-dependent correlations



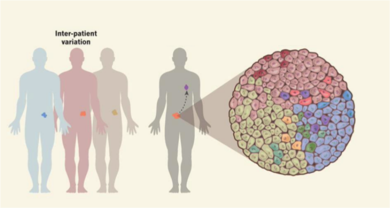
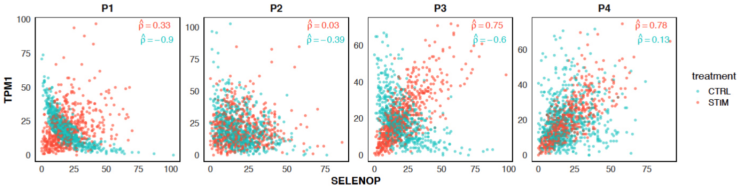
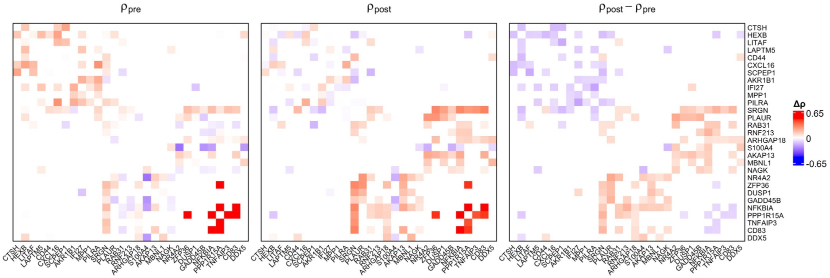


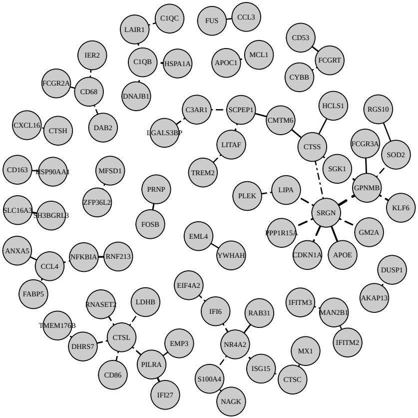
Image from Nature, Tanaka et al, 2018



- Simulation studies:
  - improved power vs. Gaussian and permutation methods
  - proper coverage and FDR control
- Real data:
  - TNBC myeloid cells (pre vs post treatment)
  - CRC CD4 vs CD8 T cells
- Identifies biologically meaningful interaction changes



scCOSMiX


C1QC
LAIR1
FUS
CCL3
CD53
IER2
C1QB
HSPA1A
APOC1
MCL1
FCGRT
FCGR2A
CD68
DNAJB1
CYBB
CXCL16
CTSH
DAB2
LGALS3BP
C3AR1
SCPEP1
CMTM6
HCLS1
RGS10
LITAF
CTSS
FCGR3A
SOD2
CD163
HSP90AA1
MFSD1
TREM2
SGK1
PRNP
PLEK
LIPA
GPNMB
ZFP36L2
KLF6
SLC16A3
SH3BGRL3
FOSB
SRGN
EML4
PPP1R15A
GM2A
ANXA5
NFKBIA
RNF213
YWHAH
CDKN1A
APOE
CCL4
DUSP1
FABP5
EIF4A2
AKAP13
RNASET2
LDHB
IFITM3
IFI6
RAB31
MAN2B1
TMEM176B
CTSL
DHRS7
EMP3
NR4A2
IFITM2
MX1
PILRA
CD86
ISG15
S100A4
CTSC
IFI27
NAGK

CNNC

Deep learning for inferring gene relationships from single-cell expression data

Ye Yuan[a] and Ziv Bar-Joseph[a,b,1]

[a]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213; and [b]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Nancy R. Zhang, University of Pennsylvania, Philadelphia, PA, and accepted by Editorial Board Member Peter J. Bickel November 12, 2019 (received for review July 7, 2019)

Several methods were developed to mine gene–gene relationships from expression data. Examples include correlation and mutual information methods for coexpression analysis, clustering and undirected graphical models for functional assignments, and directed graphical models for pathway reconstruction. Using an encoding for gene expression data, followed by deep neural networks analysis, we present a framework that can successfully address all of these diverse tasks. We show that our method, convolutional neural network for coexpression (CNNC), improves upon prior methods in tasks ranging from predicting transcription factor targets to identifying disease-related genes to causality inference. CNNC's encoding provides insights about some of the decisions it makes and their biological basis. CNNC is flexible and can easily be extended to integrate additional types of genomics data, leading to further improvements in its performance.

gene interactions | deep learning | causality inference

Several computational methods have been developed to infer relationships between genes based on gene expression data. These range from methods for inferring coexpression relationships between pairs of genes (1) to methods for inferring a biological or disease process for a gene based on other genes [either using clustering or guilt by association (2)] to causality inferences (3, 4) and pathway reconstruction methods (5). To date, each of these tasks was handled by a different computational framework. For example, gene coexpression analysis is usually performed using Pearson correlation (PC) or mutual information (MI) (6). Functional assignment of genes is often performed using clustering (7) or undirected graphical models including Markov random fields (8), while pathway reconstruction is often based on directed probabilistic graphical models (4). These methods also serve as an initial step in some of the most widely used tools for the analysis of genomics data including network inference and reconstruction approaches (3, 9, 10), methods for classification based on genes expression (11) and many more.

While successful and widely used, these methods also suffer from serious drawbacks. First, most of these methods are unsupervised. Given the large number of genes that are profiled, and the often relatively small (at least in comparison) number of samples, several genes that are determined to be coexpressed or cofunctional may only reflect chance or noise in the data (12). In addition, most of the widely used methods are symmetric, which means that each pair has only one relationship value. While this is advantageous for some applications (e.g., clustering), it may be problematic for methods that aim at inferring causality (e.g., network reconstruction tasks).

To address these issues, we developed a method, convolutional neural network for coexpression (CNNC), which provides a supervised way (that can be tailored to the condition/question of interest) to perform gene relationship inference. CNNC utilizes a representation of the input data specifically suitable for deep learning. It represents each pair of genes as an image (histogram) and uses convolutional neural networks (CNNs) to infer relationships between different expression levels encoded in the image. The network is trained with positive and negative examples for the specific domain of interest (e.g., known targets of a transcription factor [TF], known pathways for a specific biological process, known disease genes, etc.), and the output can be either binary or multinomial.

We applied CNNC using a large cohort of single-cell (SC) expression data and tested it on several inference tasks. We show that CNNC outperforms prior methods for inferring interactions (including TF–gene and protein–protein interactions), causality inference, and functional assignments (including biological processes and diseases).

Results

We developed CNNC, a general computational framework for supervised gene relationship inference (Fig. 1). CNNC is based on a CNN, which is used to analyze summarized co-occurrence histograms from pairs of genes in single-cell RNA-sequencing (scRNA-seq) data. Given a relatively small labeled set of positive pairs (with either negative or random pairs serving as negative), CNNC learns to discriminate between interacting, causal pairs, negative pairs, or any other gene relationship types that can be defined.

Learning a CNNC Model. CNNC can be trained with any expression dataset, although as with other neural network applications, the more data, the better its performance. Given expression data, we

Significance

Accurate inference of gene interactions and causality is required for pathway reconstruction, which remains a major goal for many studies. Here, we take advantage of 2 recent technological developments, single-cell RNA sequencing and deep learning to propose an encoding scheme for gene expression data. We use this encoding in a supervised framework to perform several different types of analysis using minimal assumptions. Our method, convolutional neural network for coexpression (CNNC), first transforms expression data lacking locality to an image-like object on which convolutional neural networks (CNNs) work very well. We then utilize CNNs for learning relationships between genes, causality inferences, functional assignments, and disease gene predictions. For all of these tasks, CNNC significantly outperforms all prior task-specific methods.

Author contributions: Y.Y. and Z.B.-J. designed research; Y.Y. and Z.B.-J. performed research; Y.Y. analyzed data; and Y.Y. and Z.B.-J. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. N.R.Z. is a guest editor invited by the Editorial Board.

Data deposition: The software in this paper has been deposited in GitHub, https://github.com/xiaoyeye/CNNC.

[1]To whom correspondence may be addressed. Email: zivbj@cs.cmu.edu.

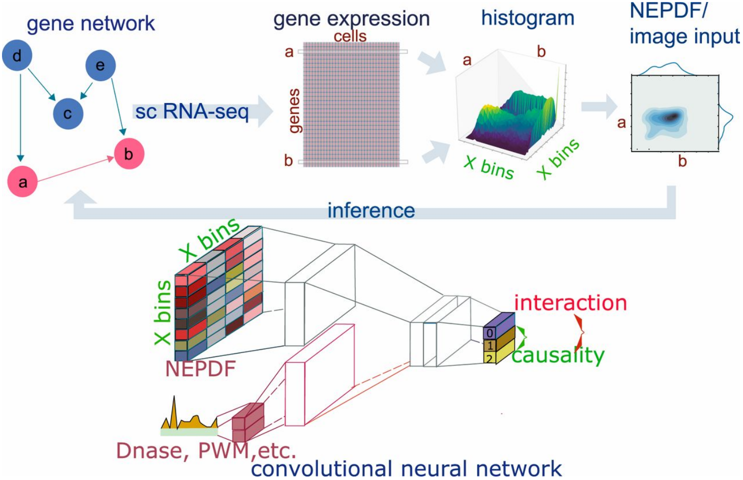
This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1911536116/-/DCSupplemental.

First published December 10, 2019.

SYSTEMS BIOLOGY

COMPUTER SCIENCES

CNNC


gene network
d
e
c
a
b
sc RNA-seq
gene expression
cells
a
b
genes
histogram
a
b
X bins
X bins
NEPDF/
image input
a
b
inference
X bins
X bins
NEPDF
interaction
0
1
2
causality
Dnase, PWM,etc.
convolutional neural network