Effective expression analysis using gene interaction matrices and convolutional neural networks

Arvind Pillai[1], Piotr Grabowski[2] and Bino John[1, *]

[1]Imaging and Data Analytics, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Waltham, US

[2]Imaging and Data Analytics, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Cambridge, UK

*To whom correspondence should be addressed. bino.john@astrazeneca.com

Abstract

Artificial intelligence recently experienced a renaissance with the advancement of convolutional neural networks (CNNs). CNNs require spatially meaningful matrices (*e.g.*, image data) with recurring patterns, limiting its applicability to high-throughput omics data. We present GIM, a simple, CNN-ready framework for omics data to detect both individual and network-level entities of biological importance. Using gene expression data, we show that GIM-CNNs can outperform comparable neural networks in performance and their design facilitates network-level interpretability. GIM-CNNs provide a means to discover novel disease-relevant factors beyond individual genes and their expression, factors that are likely missed by standard differential gene expression approaches.

Introduction

Convolutional neural network (CNN) architectures have emerged as one of the most powerful approaches in machine learning[1,2]. The use of CNN-centric architectures enables automatic inference of relevant features from image-based data and allows data scientists to leverage the rich array of methodologies in both CNNs and, more broadly, deep learning (DL). However, due to the vectorized nature of omics data, generic frameworks to analyze expression data using CNNs has been limited with the added caveat of loss of interpretability with the best performing methods. For example, DeepInsight (DI) converts gene-expression data to an image-like feature matrix using unsupervised dimensionality reduction (t-SNE, k-PCA) and image processing[3]. DI uses complex data transformations using latent features to yield the relevant image matrices for CNNs, making it a powerful tool. However, such complex transformations also make it nearly impossible to provide a biologically meaningful interpretation of the resulting dimensionality-reduced features. A probabilistically-grounded framework to predict gene relationships from single cell sequencing data using a co-expression based approach[4] that derives a normalized empirical probability distribution function (NEPDF) for CNNs was also recently developed. NEPDFs are powerful as additional inputs to CNNs to predict gene-gene relations. NEPDFs currently rely on processing of very large sets of relevant data such as single cell sequencing to build the underlying NEPDF matrices. More recently, several different CNN architectures were tested[5] to predict cancer types using RNA-Seq data from TCGA[6]. The resulting observations further substantiates previous studies on the advantages of using CNNs for gene expression. However, the end-users are restricted by the CNN-architectures and are limited to classifying cancer types as implemented in the code. Due to the restricted availability and limitations of methods available for applying CNNs on gene-expression data, we developed a novel, generalized approach, termed GIM (Gene Interaction Matrices). GIM uses a biologically inspired,

gene-interaction based data transformation on gene expression data to create an image-like feature matrix from any gene expression-based study. The transformed data can then be used with any CNN-based machine learning approach for a variety of challenging problems such as disease diagnostics and drug development. We compared the performances of a number of standard neural network architectures including GIM-CNNs on two disparate datasets[6,7] to illustrate the utility of GIM-CNNs in classification problems. We further use kidney cancer as an example to illustrate the ability of GIMs to unravel disease relevant interactions. We show that GIM-CNNs can identify important differentially regulated genes, as well as complex gene-gene links that are non-trivial to uncover using standard differential gene expression (DEG) techniques.

Methods

Detailed information on datasets, feature selection, and testing strategies are described in the supplementary material. GIM is designed to make use of high-throughput readouts of defined entities such as those of genes (*e.g.*, gene expression data) from the primary samples of interest (*e.g.*, treatment samples), and from the samples that represent the corresponding baseline biological signals, such as “control” samples. For example, given two study groups of a gene expression experiment such as treated vs control samples, we denote the corresponding gene expression data by $T \in \mathbb{R}^{n_t \times n_g}$, and $C \in \mathbb{R}^{n_c \times n_g}$, where n_t, n_c, and n_g are number of treatment replicates, number of control replicates, and genes of interest (*e.g.*, most variant genes). GIM uses the gene expression datasets to compute a square matrix (A) as the input for a CNN, comprising of a harmonic gene-gene score (diagonal and lower triangular matrix), and a relative gene-gene expression score (upper triangular matrix). Specifically, given two genes g_i and g_j, the cell value, A_{ij} for a sample is calculated as shown (equations 1-3).

$$A_{ij} = \begin{cases} \ln\left(RR\left(g_i, g_j\right)\right) & i < j \\ \ln\left(HR\left(g_i, g_j\right)\right) & i \geq j \end{cases} \quad (1)$$

$$RR\left(g_i, g_j\right) = \frac{\sum_{r=1}^{n_t}\left(T_{g_i}^r / T_{g_j}^r\right)/n_t}{\sum_{r=1}^{n_c}\left(C_{g_i}^r / C_{g_j}^r\right)/n_c + \varepsilon} \quad (2)$$

$$HR\left(g_i, g_j\right) = \frac{\sum_{r=1}^{n_t}\left[\left(2 \times T_{g_i}^r \times T_{g_j}^r\right)/\left(T_{g_i}^r + T_{g_j}^r\right)\right]/n_t}{\sum_{r=1}^{n_c}\left[\left(2 \times C_{g_i}^r \times C_{g_j}^r\right)/\left(C_{g_i}^r + C_{g_j}^r\right)\right]/n_c + \varepsilon} \quad (3)$$

Where $i, j = 1 \ldots n_g$, $\varepsilon = 1 \times 10^{-5}$, $T_{g_i}^r$ and $T_{g_j}^r$ represents value of genes g_i and g_j in the r-th treatment replicate, respectively; $C_{g_i}^r$ and $C_{g_j}^r$ represents value of genes g_i and g_j in the r-th control replicate, respectively. Note that for diagonal elements of *A* (*i.e. i=j*), the harmonic mean ratio (*HR*) scores reduces to the equivalent of fold-change values for each gene, allowing CNNs to explicitly use this important feature. *HR* scores thus capture the changes in average cumulative abundance of gene pairs between two experimental conditions. In contrast, relative ratio (RR) scores measure the changes in the relative abundance of gene pairs (*e.g.*, perturbations in inter-dependent pathways). Replacement of genes by other biological entities of interest in the equations would allow for the application of GIMs to other similar data such as proteomics and metabolomics. GIM is made available for free through the AstaZeneca R&D Github (https://github.com/AstraZeneca/GIM).

Results

We tested GIM in two different scenarios (Figure 1a): (1) a binary classification problem with matched treatment and control samples using the Open TG-Gates (TGG) data[7] , and (2) a more

intricate, multi-class classification problem on TCGA datasets with the added complexity of not having an experimentally defined, matched controls[6]. GIM was first applied to the TGG dataset of 160 compounds using a CNN (GIM-CNN) to predict if a given compound-induced gene expression perturbation, is adverse. GIM-CNN yielded the highest MCC[8] scores of 0.42 over all different neural network-based architectures tested, corresponding to an improvement of ~10-20% in performance over the other methods. Even with the use of thousands more features (7400 vs 225 genes) the next best performing method, CNN-Re[5] only yielded an MCC of 0.38. For the second test using the TCGA cancer dataset, we used GIM-CNN to predict 33 cancer types. Once again, GIM-CNN performed better than all other methods tested when using an identical feature set of 225 genes. GIM also enabled highly comparable classification accuracy to that of CNN-Re (93.10 vs 93.8 %), while using only a small proportion (3%) of the features used by CNN-Re[5]. Finally, since each GIM feature represents a pseudo measure of biologically relevant interactions, GIM-CNNs presents a direct route to discover potentially important regulators, interactions, and networks (Fig. 1C). For example, the top 30 important gene interactions for Kidney Renal Clear Cell Carcinoma naturally, and perhaps surprisingly, fall into a well-connected network with UPK1B, GPX2, and AQP4 representing the top three hubs, all of which have been linked to renal cancer. For instance, the top connected gene, UPK1B has recently emerged as an essential gene for renal urothelium function[9], and is a known promoter of bladder metastasis[10]. GPX2, a renal cancer-promoting gene, manifests high expression in specific clinical sub-types only[11], highlighting the feature of GIM-CNNs in extracting meaningful links that are challenging for typical DEG-based approaches. Finally, AQP4, belonging to the aquaporin family, has an established role in cancer biology[12] with clinical links to bladder cancer[13,14], suggestive of an underappreciated role for AQP4 in renal cancer biology.

In summary, GIM-CNNs presents several other advantages over other machine learning methods for gene expression data analysis, simultaneously enabling network-level biological discoveries and strong predictive performance while using modest number of features. In contrast to non-DL methods, GIM allows for leveraging the rich and complex information hidden in a high-dimensional gene expression dataset. With respect to other DL methods such as DI for gene expression, GIM also provides the advantage of having a deterministic feature encoding step. These attributes of GIM should improve reproducibility and interpretability in machine learning workflows using CNNs. The proposed framework can be easily also applied to similar high-dimensional vectorized data including but not limited to epigenome profiling (*e.g.,* ATAC-Seq, ChIP-Seq), and proteome profiling (*e.g.*, mass-spec proteomics), and metabolomics. Since CNN's can make use of high-dimensional tensors, GIM framework could also enable additional data about gene-gene relationships (*e.g.,* PPI, proteomics data) in future studies. GIM may also prove useful in leveraging the powerful transfer learning approach (*e.g.,* ResNet[15]) to address machine learning problems where large training datasets are not available. For instance, GIM-CNNs could be pre-trained on large datasets such as TCGA and single cell sequencing, and subsequently fine-tuned with smaller datasets of interest for predictive as well as hypothesis generation needs.

Acknowledgements

Sincere thanks to AstraZeneca (AZ) internal reviewers, Nigel Greene (AZ), and Mishal Patel (AZ).

Conflict of Interest: none declared.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

2. Anwar, S. M. *et al.* Medical Image Analysis using Convolutional Neural Networks: A Review. *J. Med. Syst.* **42**, 226 (2018).

3. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A. & Tsunoda, T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* **9**, 1–7 (2019).

4. Yuan, Y. & Bar-Joseph, Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 27151–27158 (2019).

5. Mostavi, M., Chiu, Y. C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med. Genomics* **13**, 1–13 (2020).

6. Network, C. G. A. R. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

7. Igarashi, Y. *et al.* Open TG-GATEs: A large-scale toxicogenomics database. *Nucleic Acids Res.* **43**, D921–D927 (2015).

8. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).

9. Carpenter, A. R. & McHugh, K. M. Role of renal urothelium in the development and progression of kidney disease. *Pediatr. Nephrol.* **32**, 557–564 (2017).

10. Wang, F. H., Ma, X. J., Xu, D. & Luo, J. UPK1B promotes the invasion and metastasis of bladder cancer via regulating the Wnt/β-catenin pathway. *Eur. Rev. Med. Pharmacol. Sci.* **22**, 5471–5480 (2018).

11. Naiki, T. *et al.* GPX2 promotes development of bladder cancer with squamous cell differentiation through the control of apoptosis. *Oncotarget* **9**, 15847–15859 (2018).

12. Papadopoulos, M. C. & Saadoun, S. Key roles of aquaporins in tumor biology. *Biochim. Biophys. Acta - Biomembr.* **1848**, 2576–2583 (2015).

13. Figueroa, M. *et al.* Paraneoplastic Neuromyelitis Optica Spectrum Disorder

Associated With Metastatic Carcinoid Expressing Aquaporin-4. *JAMA Neurol.* **71**, 495–498 (2014).

14. Yi, S. & Park, H. A rare case of aquaporin-4-antibody-positive neuromyelitis optica associated with bladder cancer. *Mult. Scler. Relat. Disord.* **38**, 101499 (2020).

15. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

A

Model	TGG[f]	MCC	TCGA[f]	Accuracy
ANN	225	0.37 ± 0.05	225	87.85 ± 0.54
CNN-Re	225	0.32 ± 0.13	225	88.43 ± 0.90
CNN-Re	7400	0.38 ± 0.10	7100	93.82 ± 0.46
DI	225	0.37 ± 0.16	225	92.18 ± 0.30
DI	7473	0.34 ± 0.04	24062	94.18 ± 0.40
GIM	225	0.42 ± 0.07	225	93.10 ± 0.16

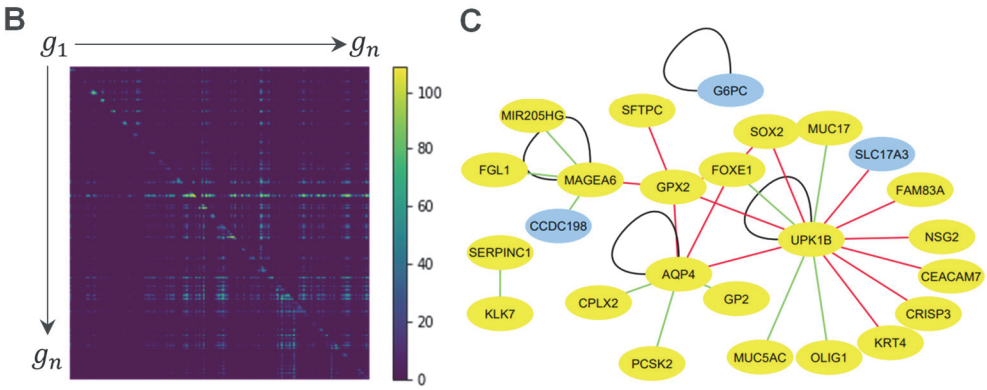

Figure 1: **A)** Performance comparisons of various methods on TGG and TCGA datasets (See supplement for details). ANN: Artificial neural network, CNN-Re: CNN-reshape, DI: DeepInsight, MCC: Matthew's correlation coefficient; Accuracy: Categorical Accuracy. TGG[f] and TCGA[f] represent number of features used for the TGG, and TCGA datasets, respectively. **B)** Visualization of the GIM feature importance map for Kidney Renal Clear Cell Carcinoma (KIRC) using GRAD-CAM. Brighter color corresponds to most important GIM features (gene pairs) for KIRC classification. **C)** Top 30 most important features (gene pairs) extracted from the feature importance map (Fig 1b), spanning 25 unique genes. Edge colors correspond to types of feature interactions. Green: harmonic mean score (HR, lower triangle of the matrix), black: gene log2 fold-change (HR scores at diagonal of the matrix), red: relative ratio score (RR, upper triangle of the matrix). Known KIRC biomarkers are marked as blue nodes.