Learning Functional Embedding of Genes Governed by Pair-wised Labels

Jingjun Cao
Chinese Academy of Agricultural Sciences
Beijing, China
Email: caojingjun@caas.cn

Zhenglin Wu
International School
Beijing University of Posts and Telecommunications
Beijing, China
Email: wuzhenglin@bupt.edu.cn

Wenting Ye
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China
Email: wenting_ye@bupt.edu.cn

Haohan Wang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
Email: whaohan@acm.org

***Abstract*—In this work, we build a deep neural network architecture which learns a compact numerical representation of genes supervised by numerous sources of pair-wise information, including Protein-Protein Interaction information and Gene Ontology information. We introduce a new network architecture which can process gene expression data and generate the representation of individual genes while governed by pair-wise information. The learnt representation is aimed to be further used for research of bioinformatics on relevant tasks, and even beyond the information sources from embedding learnt. Within this paper, we evaluate the representation on Protein-Protein Interaction task, and it shows a result which is better than learnt representation from traditional dimension reduction and feature selection methods.**

***keywords*—** Representation Learning, Deep Learning, Computational Biology

I. Introduction

Computational biology research lies on the intersection between molecular biology and machine learning. In recent years, a lot of novel solutions have been proposed for biological research and medical application. For example, [1] improved clustering algorithms to differentiate breast cancer patients and normal people from gene expression data. [2] developed a method to use brain wave to help monitor the students' performance during online education. [3] proposed novel models for the genetic studies. [4] introduced a time series model to monitor the process of mental states. Nowadays, we are expecting to see a blooming of the research area between machine learning and biology. The connection of these two domains can be built by encoding biological processes into numerical representations and feeding the representations into machine learning algorithms. An effective representation of biological processes consequently becomes more and more essential. For a gene, such a representation can be either biological experimental results, Microarray/RNAseq expressions, or sequence information of either amino acid sequence or nucleotide sequence. For example, Protein-Protein Interaction (PPI) prediction is a field that utilizes state-of-art machine learning models and algorithms based on biological data and aims to predict the new PPI pairs statistically. The biological data could be sequence structure ([5]), gene expression data ([6]), or some high level features that extracted from biological experiments ([7]).

However, these data sources may come with problems. Experimental results or gene expression data can hardly represent the corresponding gene comprehensively. Sequence information can hardly be interpreted efficiently. Here in this work, we build a network to learn a vector representation of genes that could encode the relevant information from gene expression and some other experimental results. We believe that the representation can further be used for research of bioinformatics on relevant tasks, or even beyond the information sources from the embedding learned. Another advantage of our work is that we could easily incorporate pair-wise or even group-wise information from genes as the sources of information while learning the representations. However, the group information cannot easily be represented with existing methods for a single gene.

II. Related Work

Several attempts have been made to learn useful representations from gene expression data, although most of them are tailored for specific learning tasks.

As one of the first attempts, [8] learned representation from gene expression data for cluster structure. [9] learned expression by applying ICA to gene expression data and used the result as features for a discriminative classifier for the task of tumor classification. [10] achieved considerable success on feature selection methods applied to genomic microarray data. The feature selection method that they proposed partially focuses on the redundant features, which is a popular problem in the microarray analysis domain. Also, focusing on removing redundant features, [11] proposed a minimum redundancy

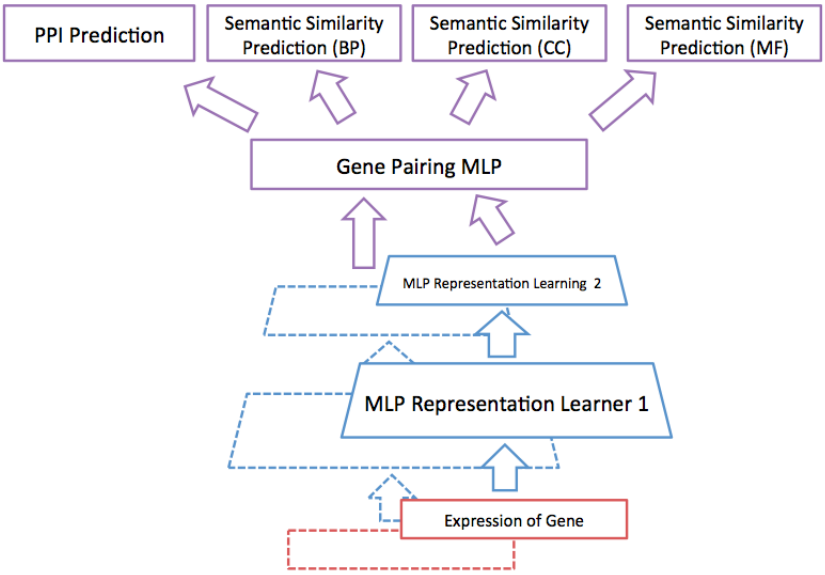

Fig. 1. An illustration of multi-task network architecture.

maximum relevance (MRMR) feature selection framework for gene expression analysis about the problem of cancer classification. The aforementioned method achieves consistent success for many different classifiers.

Over the last few years, deep learning methods have emerged as the alternatives to traditional machine learning methods, and deliver state-of-the-art performance for many learning tasks [12], [13]. [14] used deep architectures, pre-trained in an unsupervised manner using denoising auto-encoders, as a preprocessing step to regenerate gene expression time series data for two different data sets and to show promising improvement in performance. [15] took a step further and showed convolutional kernel of deep learning as an effective tool to calculate correlations of two sequences and further used such technique to build a more powerful deep learning architecture to learn the representation of paired-wise information.

In this paper, we further extend the previous work by introducing a new network architecture that can learn the representation of individual genes governed by pair-wise information.

III. Method

In this section, we first introduce the new network architecture which can learn the representation of individual genes while governed by pair-wise information, then we will briefly introduce the techniques we use in parameter learning.

A. Model

Figure 1. shows the basic architecture of our network. Each blue component represents the network for learning representations. Each purple component represents a network for solving a single task. Red components are input data. We introduce the network architecture bottom up [16].

a) Representation Learner: The bottom two layers are traditional neural networks that serve as representation learner. The representation of genes will be achieved from the top of the second representation learner.

b) Gene Pairing Layer: This layer pairs the representation generated from below layers and concatenates the representation for pair-wise tasks. In order to guarantee the genuineness of learned representation, we want to reduce the complexity of network architecture above representation learners. Therefore, weights of this gene pairing layer only server as the simple concatenation and do not get tuned while training.

c) PPI Prediction: This task solver takes the output of the previous layer as the input vector to predict PPI. Mean Squared Errors will be backpropagated to previous layers for updating weights.

d) Semantic Similarity Prediction: Semantic similarity prediction solves the task for predicting how similar these two genes are, in terms of three different GO Sub-Ontologies. The similarity is defined with five different approaches, proposed by [17], [18], [19], [20], [21]. Therefore, we have 15 semantic similarity prediction task solvers altogether. The similarity is defined as a continuous value from 0 to 1, so each of these task solvers is to solve a regression problem.

B. Parameter Learning

Based on the introduction above, the parameters of the model are tuned while solving the following optimization problem:

$$e = \frac{1}{n}\sum_{k=1}^{n}\{(\hat{y}_k^i - y_k^i)^2 + \sum_{so=1}^{3}\sum_{sim=1}^{5}(\hat{y}_k^{so,sim} - y_k^{so,sim})^2\} + \lambda_1||W||_1 + \lambda_2||W||_2^2$$

where y^i stands for label of prediction, $y^{so,sim}$ stands for similarity of Sub-Ontology so, with similarity defined as Method sim, $\hat{}$ stands for predicted labels. λ and W stands for L_1 and L_2 regularization.

The parameters of the model are learned with ADMM [22], by which we split our multiple tasks into two sets: PPI prediction task and semantic similarity prediction tasks. By splitting the tasks, we could monitor the influences of different sets of tasks over representation learning.

At each iteration, updates of parameters are given by traditional backpropagation, with several techniques plugged in, including dropout [23], layer-wise learning rate and layer-wise regularization weight. Lower layers get larger learning rates, and so that the whole model could be efficiently updated. Lower layers get more regularized due to the high variance nature of data.

C. Representation Extraction

We use the parameters trained in the previous section for the new network architecture. Then the new neural network that can process gene expression data generates the representation of individual genes while governed by pair-wise information. The learned representation can be used for research of bioinformatics on relevant tasks, such as PPI and so on.

TABLE I
RESULTS DURING TRAINING

Dimension	PPI	PTM	BP	CC	MF
20	0.4974	0.5161	0.071548	0.105373	0.105426
40	0.4996	0.5161	0.0707225	0.105053	0.104987
60	0.4674	0.4434	0.0699832	0.104132	0.104411
80	0.4936	0.5038	0.0719199	0.097035	0.095230
100	0.4546	0.4542	0.0693261	0.104177	0.104771

IV. Experimental Results

In this section, we will validate the performance of our proposed representation extractor on different perspectives. At first, we will show that the performance of model during training process demonstrates that it indeed merges different sources of knowledge into one vector representation. Then, we will present how these representations can be used to other biomedical tasks. We will start from the tasks that offer the information for the training of our model, then we will show that how the representation can behave on novel tasks which are not used to train the model.

A. Representation Learning

We set the dimension as what we are interested in, D = 20;40;60;80;100 and then learn a compact representation of dimension D. Then we feed learnt representation into traditional classifiers. We assign 1000 to epoch, the results are showed in Table 1.

B. Usage of Representation

1) Using Representation of Knowledge: We feed learnt representation into traditional classifiers, and intentionally select GaussianNB, kNeighborsClassifier, LogisticRegression and SVC. The results are showed in Table 2, where we compared our learnt representation information with origin data.

2) Using Representation of Pair-Wise Knowledge: We feed the learnt pair-wise representation into traditional classifiers, we also select GaussianNB, kNeighborsClassifier, LogisticRegression and SVC. The results are showed in Table 3, where we compared our learnt pair-wise representation information with origin data.

V. Conclusion

In this paper, we aim at learning a compact knowledge representation of the massive gene expression data supervised by numerous sources of pair-wise information, thus the compact representation can be used further with lighter computation units with almost as much information as the original intact features.

We worked on reducing the dimension of gene expression data supervised by numerous sources of pair-wise information for a specific task: to get a compact and informative representation for Protein-Protein Interactions.

In this work, we build a network to learn a vector representation of genes that could encode the relevant information from gene expression and some other experimental results.

During the evaluation phase, we evaluated our model with classification accuracy for Protein-Protein Interaction compared with learned representation information through the new neural network with origin data. Our experiments have shown that our proposed method works best in two extreme scenarios, 1)when setting dimension to 80 or 100. 2)to learn a representation that can retain as much information as possible.

Our results on the mean squared error of bp, cc and of seem all right. Training on bigger datasets will reduce the error rate even further. Using supervised learning in the Protein-Protein Interaction predicting could also be beneficial. Future work will focus on reducing the classification error of PTM and PPI.

Looking into future, there are many attractive directions that we can explore. We could easily incorporate pair-wise or even group-wise information from genes as sources of information while learning the representations. However, the group information cannot easily be represented with existing methods for a single gene, but we would like to explore some further relevant techniques in this direction.

In terms of deep learning techniques, in the future, we would like to explore more localized method [24], [25], and transfer deep learning method [26]. Recently, Select-Additive Learning [27] has become an important and promising method that will surly improve the performance in many different deep learning applications.

TABLE II
USAGE OF LEARNT REPRESENTATION

Data	GaussianNB	kNeighborsClassifier	LogisticRegression	SVC
origin data	0.0881	0.4689	0.3968	0.4661
learnt representation D=20	0.5202	0.4620	0.5013	0.5201
learnt representation D=40	0.5056	0.4445	0.4899	0.4910
learnt representation D=60	0.5097	0.4625	0.4918	0.5010
learnt representation D=80	0.3856	0.4510	0.5234	0.5320
learnt representation D=100	0.4061	0.4569	0.5331	0.5404

TABLE III
USAGE OF LEARNT PAIR-WISE REPRESENTATION

Data	GaussianNB	kNeighborsClassifier	LogisticRegression	SVC
origin data	0.5051	0.4884	0.5229	0.5009
learnt representation D=20	0.4740	0.4901	0.4741	0.4851
learnt representation D=40	0.5007	0.4829	0.5032	0.5087
learnt representation D=60	0.4794	0.4763	0.4955	0.4921
learnt representation D=80	0.5303	0.5404	0.5302	0.5404
learnt representation D=100	0.5470	0.5014	0.5186	0.5208

ACKNOWLEDGMENT

The authors would like to thank Ben Lengerich from Carnegie Mellon University for the suggestions of the presentation of this paper.

REFERENCES

[1] Haohan Wang, Lixiang Li, Xi Yang, and Chong Lian. A novel chaotic ant swarm based clustering algorithm for clinical prediction. *Research Journal of Applied Sciences, Engineering and Technology*, 4(20):3981–3988, 2012.

[2] Haohan Wang, Yiwei Li, Xiaobo Hu, Yucong Yang, Zhu Meng, and Kai-min Chang. Using eeg to improve massive open online courses feedback interaction. In *AIED Workshops*, 2013.

[3] Haohan Wang and Jingkang Yang. Multiple confounders correction with regularized linear mixed effect models, with application in biological processes. *bioRxiv*, page 089052, 2016.

[4] Jingkang Yang, Haohan Wang, Jun Zhu, and Eric P Xing. Sedmid for confusion detection: Uncovering mind state from time series brain wave data. *arXiv preprint arXiv:1611.10252*, 2016.

[5] Joel R Bock and David A Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.

[6] Nitin Bhardwaj and Hui Lu. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.

[7] Haohan Wang and Madhavi K Ganapathiraju. Evaluation of protein-protein interaction predictors with noisy partially labeled data sets. *arXiv preprint arXiv:1509.05742*, 2015.

[8] Ka Yee Yeung and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[9] De-Shuang Huang and Chun-Hou Zheng. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*, 22(15):1855–1862, 2006.

[10] Eric P Xing, Michael I Jordan, Richard M Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608. Citeseer, 2001.

[11] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[12] Haohan Wang and Bhiksha Raj. A survey: Time travel in deep learning space: An introduction to deep learning models and how deep learning models evolved from the initial ideas. *arXiv preprint arXiv:1510.04781*, 2015.

[13] Haohan Wang and Bhiksha Raj. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*, 2017.

[14] Aman Gupta, Haohan Wang, and Madhavi Ganapathiraju. Learning structure in gene expression data using deep architectures, with an application to gene clustering. *bioRxiv*, page 031906, 2015.

[15] Haohan Wang, Aman Gupta, and Xu Ming. Extracting compact representation of knowledge from gene expression data for protein-protein interaction. *International journal of data mining and bioinformatics*, 2017.

[16] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, volume 2, page 6, 2015.

[17] Philip Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130, 1999.

[18] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[19] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[20] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7(1):302, 2006.

[21] James Z Wang, Zhidian Du, Rapeeporn Payattakool, S Yu Philip, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[22] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[24] Jiu Lin Luo, Hao Jing Luo, Ai Min Li, and Hao Han Wang. Localized model to segmentally estimate miles per gallon (mpg) for equipment engines. In *Applied Mechanics and Materials*, volume 556, pages 1069–1074. Trans Tech Publ, 2014.

[25] Ruhui Shen, Jialiang Shen, Yuhong Li, and Haohan Wang. Predicting usefulness of yelp reviews with localized linear regression models. In *Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on*, pages 189–192. IEEE, 2016.

[26] Seungwhan Moon, Suyoun Kim, and Haohan Wang. Multimodal transfer deep learning with applications in audio-visual recognition. *arXiv preprint arXiv:1412.3121*, 2014.

[27] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*, 2016.