SCIENCE CHINA
Information Sciences

• LETTER •

September 2019, Vol. 62 199104:1–199104:3
https://doi.org/10.1007/s11432-018-9704-7

DGEPN-GCEN2V: a new framework for mining GGI and its application in biomarker detection

Jinyin CHEN, Haibin ZHENG, Hui XIONG, Yangyang WU, Xiang LIN,
Shiyan YING & Qi XUAN*

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023*, China*

Received 20 July 2018/Revised 5 September 2018/Accepted 31 October 2018/Published online 29 July 2019

Citation Chen J Y, Zheng H B, Xiong H, et al. DGEPN-GCEN2V: a new framework for mining GGI and its application in biomarker detection. Sci China Inf Sci, 2019, 62(9): 199104, https://doi.org/10.1007/s11432-018-9704-7

Dear editor,

Deep learning is widely applied in various fields [1, 2] recently. In the field of bioinformatics, deep learning also attracts many attentions of researchers. Danaee et al. [3] extracted depth functional features from high-dimensional gene expression profile by stacked denoising automatic encoder and identified a set of highly interacting genes for cancer biomarkers detection. Chen et al. [4] proposed a novel spectral clustering with automatic cluster number determination, which provids new idea for cancer subtypes detection. Sedano et al. [5] developed shape-based clustering model by using time pattern of gene expression values. Singh et al. [6] proposed cascaded feature selection and stack sparse automatic encoders to learn advanced features. Liang et al. [7] proposed a novel learning model for multi-peak deep belief network, which provids effective guidance for individualized cancer treatment.

In addition, Xie et al. [8] predicted a gene expression of variant genotype based on the deep learning regression model of multi-layer perceptron and stacked denoising automatic encoder. Chen et al. [9] designed a deep learning method (D-GEX) that fully captures non-linear correlation between gene expressions, which reduces the cost of gene expression profiling inference.

A new framework to mine GGI (gene-to-gene interactions) and detect biomarkers is designed. The overall framework is shown in Figure 1. First, a three-layer model is constructed to select optimal genes with critical information and DGEPN (deep gene expression prediction network) is proposed to mine GGSI (gene-to-gene sensitivity information). Then, GCEN (gene co-expression network) is constructed based on optimal genes, which will be transformed into vectors via feature learning algorithm. Finally, for the better discovery of critical gene modules and hub genes, CCSD (cluster center self-determination) is applied to search gene subsets or modules with high compactness.

Layered gene selection model. In order to better mine the depth information contained in gene expression data, the following selection strategy is applied. The genes with differential expression, which are more relevant to sample classification and highly sensitive to the other gene expression, are retained for further study. While filtering out redundant information to simplify gene sets in gene pool, the key information of gene expression data is preserved as much as possible.

The genes with differential expression in different sample sets are selected as the 1st gene pool.

In order to obtain better GGSI, DGEPN based on the deep learning method with multi-layer neural network is proposed for extracting the expression correlation between genes. The schematic diagram of neural network architecture is shown in the second layer part of Figure 1. Considering the high-dimensional characteristics of gene expression profiles, the main body of neural network consists

* Corresponding author (email: xuanqi@zjut.edu.cn)

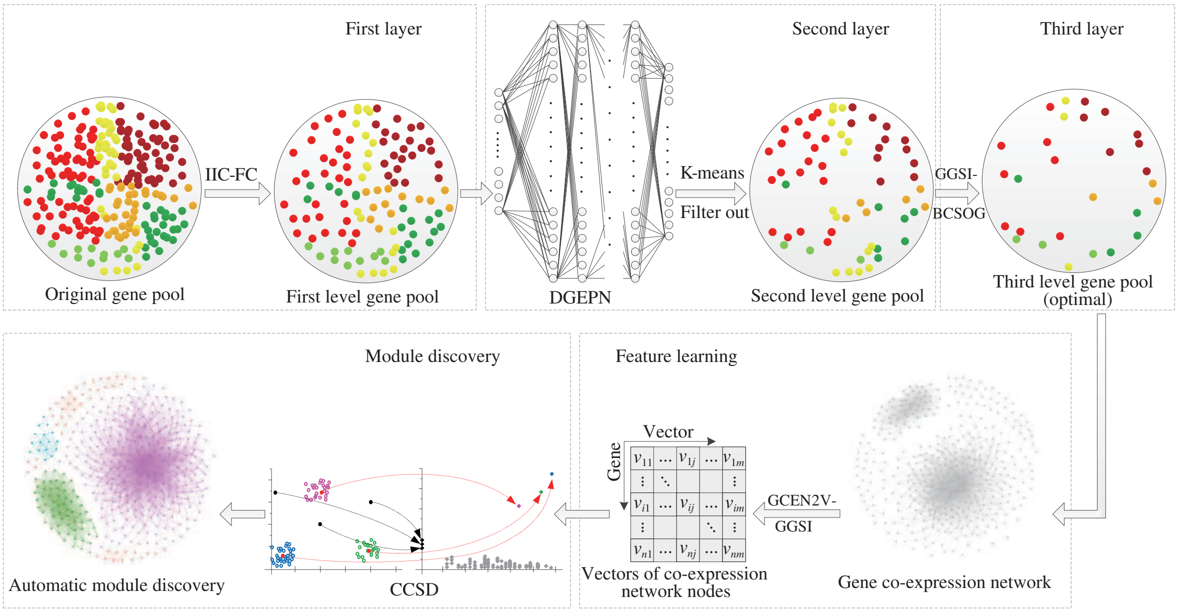

Figure 1 (Color online) The framework of the proposed method for gene expression profile study.

of fully connected modules, which can better extract sensitivity information between genes.

Assuming that the number of samples is N_{sam}, the gene dimension of each sample is M_{gene}, where M_{in} and M_{out} are the numbers of neurons in input and output layers of neural network, respectively. And each sample can be represented as $\{x_i^k|_{i=1}^{M_{\text{in}}}, y_j^k|_{j=1}^{M_{\text{out}}}\}|_{k=1}^{N_{\text{sam}}}$. The sensitivity information of the i-th input gene to j-th output gene is defined as

$$S_{ij} = \frac{1}{N_{\text{sam}}} \sum_{k=1}^{N_{\text{sam}}} \left| \frac{\partial y_j^k}{\partial x_i^k} \right|, \tag{1}$$

where $\partial y_j^k / \partial x_i^k$ represents the derivative of the j-th output neuron to the i-th input neuron. The normalized sensitivity information index is defined as follows:

$$\text{GGSI}(x_i) = \frac{\sum_j^{M_{\text{out}}} S_{ij}}{\max\{\sum_j^{M_{\text{out}}} S_{ij}|_{i=1,2,\ldots,M_{\text{in}}}\}}. \tag{2}$$

The mean square error at each output unit is defined by the loss function as follows:

$$\text{Loss} = \frac{1}{N_{\text{sam}}} \sum_{k=1}^{N_{\text{sam}}} \left[\frac{1}{M_{\text{out}}} \sum_{j=1}^{M_{\text{out}}} (y_j^k - \hat{y}_j^k)^2 \right]. \tag{3}$$

In order to avoid filtering out critical genes in the process of gene selection, the 3rd layer gene pool for co-expressing network construction is selected from the 2nd gene pool based on binary cuckoo search (BCS) algorithm.

Feature learning of GCEN via GGSI. GCEN analysis algorithm, as an efficient and accurate biological data mining tool, can identify gene modules with high co-expression trend. The main idea of network feature learning based on node2vec technology is to transform gene node feature learning of GCEN into optimizing the "possibility" of objective function, which can retain its neighbors' information.

CCSD for module discovery of GCEN. The density matrix and distance matrix are obtained by calculating the density of each gene and the minimum distance to genes with higher density. Then we draw the decision graph, marked as $\rho - \delta$, with the abscissa being density and ordinate being distance. According to the CCSD algorithm based on density peak search, the genes with higher density and relatively large distance on decision graph are selected as cluster centers. According to the characteristics of decision graph with data-intensiveness and huge data volume, a new automatic determination method of cluster centers based on plane fitting and residual analysis was designed, which combines single and multiple linear regression analysis. More detailed description about method is shown in Appendix A.

Results and discussion. Four gene expression data sets from GEO were used in the experiments, including GDS3837, GDS2771, GDS3257, and GDS2373.

In order to clearly present the proposed method, the results of GDS3837 are introduced in detail, which is microarray gene expression data of pairwise tumor and adjacent normal lung tissue specimens obtained from nonsmoking female non-small cell lung carcinoma patients.

In GDS3837, the 1st gene pool was established based on differential expression level, and the 2nd pool was screened according to GGSI score. Owing to the complexity and uncertainty of biological experiments, the quality of each gene pool is verified. When verifying the classification information abundance in the 2nd gene pool, principal component analysis (PCA) is used to extract features and support vector machine (SVM) is used for sample classification. The classification accuracy values of the 2nd gene pool obtained by different methods are basically stable, which further reflects the good performance of critical genes extracted by layered gene selection model. Finally, the 3rd gene pool was optimized via BCS. Because of uncertainty of optimization algorithm, the experiment of the 3rd gene pool construction will be repeated several times. However, the GGSI value of optimal genes selected by BCS is not the largest. The results show that only four tumor samples and one normal sample are misclassified based on optimal gene pool.

We can find that the key classification information is well preserved during gene pool establishment. Dynamic cut tree and CCSD algorithm are applied to module discovery, leading to different results. The number of modules obtained by dynamic cut tree is five, while CCSD discovered six modules. And we take GDS3837 dataset as an example of enrichment analysis about GO-terms and pathway. The most significant of the top 20 GO-terms contain negative regulation of apoptotic process, signal transduction and apoptotic process. And the most significant of the first 20 pathways from KEGG database contain pathways in cancer, proteoglycans in cancer, and transcriptional misregulation in cancer. According to the degree of enrichment about GO-terms and pathway, it can be drawn that the genes and modules obtained in this study are effective biomarkers.

Furthermore, the covariance of co-expression module discovered in the proposed method and dynamic cut algorithm is very high, which indicates that our method has good bioinformatics interpretability. Taking GDS3837 as an example, the structures of module 2 and module 4 are more compact, which have more potential research value. It can be observed that different modules have high discrimination, which indicates that connections inside modules are denser than those between modules. At the same time, the gene SEMA5A, which can be used as a biomarker for cancer detection, was also found in this experiment as potential analysis target. And the gene SEMA5A is very close to cluster center of module 2, which reflects the effectiveness of the proposed method.

The overlapped rate of gene biomarkers between baseline methods and DGEPN-GCEN2V is high, where "critical gene" represents the number of valuable genes in the references, "TOP-X" represents the number of X-genes of high potential value identified by DGEPN-GCEN2V, and "overlapped rate" is the ratio of overlapped genes to $\min\{\text{critical gene, X}\}$. The final results show good performance of the proposed method in different gene expression data, which can discover more compact co-expression modules with biological significance and more comprehensive gene biomarkers of potential value. More specific description about experimental performance is shown in Appendix B.

Acknowledgements This work was partially supported by Zhejiang Natural Science Foundation (Grant No. LY19F020025), National Natural Science Foundation of China (Grant Nos. 61502423, 61572439), Zhejiang University Open Fund (Grant No. 2018KFJJ07), Zhejiang Science and Technology Plan Project (Grant Nos. LGF18F030009, 2017C33149), and Zhejiang Outstanding Youth Fund (Grant No. LR19F030001).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Xuan Q, Fang B W, Liu Y, et al. Automatic pearl classification machine based on a multistream convolutional neural network. IEEE Trans Ind Electron, 2018, 65: 6538–6547

2 Qu W, Wang D L, Feng S, et al. A novel cross-modal hashing algorithm based on multimodal deep learning. Sci China Inf Sci, 2017, 60: 092104

3 Danaee P, Ghaeini R, Hendrix D A. A deep learning approach for cancer detection and relevant gene identification. In: Proceedings of Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, Hawaii, 2017. 219–229

4 Chen J Y, Wu Y Y, Lin X, et al. DOE-AND-SCA: a novel SCA based on DNN with optimal eigenvectors and automatic cluster number determination. IEEE Access, 2018, 6: 20764–20778

5 Chira C, Sedano J, Villar J R, et al. Gene clustering for time-series microarray with production outputs. Soft Comput, 2016, 20: 4301–4312

6 Singh V, Baranwal N, Sevakula R K, et al. Layerwise feature selection in stacked sparse auto-encoder for tumor type prediction. In: Proceedings of Bioinformatics and Biomedicine, Shenzhen, 2016. 1542–1548

7 Liang M X, Li Z Z, Chen T, et al. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Trans Comput Biol Bioinf, 2015, 12: 928–937

8 Xie R, Quitadamo A, Cheng J L, et al. A predictive model of gene expression using a deep learning framework. In: Proceedings of Bioinformatics and Biomedicine, Shenzhen, 2016. 676–681

9 Chen Y F, Li Y, Narayan R, et al. Gene expression inference with deep learning. Bioinformatics, 2016, 32: 1832–1839