OXFORD

G3 Genes | Genomes | Genetics

G3, 2022, **12(11),** jkac243
https://doi.org/10.1093/g3journal/jkac243
Advance Access Publication Date: 19 September 2022
Investigation

NESM: a network embedding method for tumor stratification by integrating multi-omics data

Feng Li ,[1] Zhensheng Sun ,[1] Jin-Xing Liu ,[1] Junliang Shang ,[1] Lingyun Dai ,[1] Xikui Liu ,[2] Yan Li ,[2,*]

[1]School of Computer Science, Qufu Normal University, Rizhao 276826, China,
[2]Department of Electrical Engineering and Information Technology, Shandong University of Science and Technology, Jinan, Shandong 250031, China

*Corresponding author: Department of Electrical Engineering and Information Technology, Shandong University of Science and Technology, Jinan, Shandong 250031, China. Email: liyan@sdkd.net.cn

Abstract

Tumor stratification plays an important role in cancer diagnosis and individualized treatment. Recent developments in high-throughput sequencing technologies have produced huge amounts of multi-omics data, making it possible to stratify cancer types using multiple molecular datasets. We introduce a Network Embedding method for tumor Stratification by integrating Multi-omics data. Network Embedding method for tumor Stratification by integrating Multi-omics pregroup the samples, integrate the gene features and somatic mutation corresponding to cancer types within each group to construct patient features, and then integrate all groups to obtain comprehensive patient information. The gene features contain network topology information, because it is extracted by integrating deoxyribonucleic acid methylation, messenger ribonucleic acid expression data, and protein–protein interactions through network embedding method. On the one hand, a supervised learning method Light Gradient Boosting Machine is used to classify cancer types based on patient features. When compared with other 3 methods, Network Embedding method for tumor Stratification by integrating Multi-omics has the highest AUC in most cancer types. The average AUC for stratifying cancer types is 0.91, indicating that the patient features extracted by Network Embedding method for tumor Stratification by integrating Multi-omics are effective for tumor stratification. On the other hand, an unsupervised clustering algorithm Density-Based Spatial Clustering of Applications with Noise is utilized to divide single cancer subtypes. The vast majority of the subtypes identified by Network Embedding method for tumor Stratification by integrating Multi-omics are significantly associated with patient survival.

Keywords: cancer subtype; multi-omics; pan-cancer; embedding network

Introduction

Cancer is generally due to a variety of factors under the occurrence of somatic variation, which can lead to the cell growth of abnormal regulation and the formation of abnormal lesions (Zhong *et al.* 2021). There are differences in cellular morphology and tissue structure between neoplastic and normal tissues. Benign neoplasms are often present relatively low atypia and same with the normal tissues from which they originate, while malignant neoplasms are present relatively high atypia (Liu *et al.* 2021). In recent years, the development of next-generation sequencing technologies and the development of several multicenter cancer exome/genome projects, The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (Chang *et al.* 2013; Jennings and Hudson 2016), provide a large amount of omics data, such as gene expression data, copy number variation, burst, and deoxyribonucleic acid (DNA) methylation data. Thus, the rapid accumulation of multi-omics tumor data has brought new opportunities and challenges to study systems biology from multilevel (Ruan *et al.* 2019).

The stratification of tumors into clinical and biological subtypes benefits precision oncology. For example, an entropy-based consensus clustering (ECC) method (Liu *et al.* 2017) for patient stratification fuses multiple base partitions into a consensus model using an entropy-based utility function. Importantly, numerous data-driven approaches for classifying cancers based on diverse clinical data have been proposed, such as multiple gene classifiers for breast cancer prognosis based on gene expression profiles (Reis-Filho and Pusztai 2011), neural network-based survival prediction for different breast cancer subtypes by combining with clinical information including tumor size and axillary lymph node status (Lundin *et al.* 1999), and skin cancer classification based on imaging data using deep neural network algorithms (Esteva *et al.* 2017).

Integrating multiplatform molecular data, such as gene expression data, miRNA expression data, and DNA methylation data, can effectively identify cancer subtypes (Liang *et al.* 2021), which has been proven to be more powerful than a single data type (Wang *et al.* 2014). Multiple strategies have posited for the integration of multiple sets of data. One strategy is to analyze each data type individually before integrating multiple sets of data (Shen *et al.* 2009; Mo *et al.* 2013), but this strategy cannot capture relationships between same-origin data. Scluster (Ge *et al.* 2017) and joint and individual variation explained (Lock *et al.* 2013) can capture the association information both between and within

Received: June 28, 2022. **Accepted:** August 30, 2022

multiple data simultaneously, but they are sensitive to feature selection. To effectively extract shared and complementary information concealed in a variety of biological data types, more systematic and integrated methodologies are required (Zhao and Yan 2020). However, biomolecular networks contain many different layers and different organizational forms in biological systems, which have been widely used in cancer research (Leiserson *et al.* 2015; Horn *et al.* 2018; Liu *et al.* 2020; Zhang and Wang 2022). Therefore, network-based strategy is an effective method to analyze and integrate multi-omics data (Ma'ayan 2011; Zhao *et al.* 2015; Zhu *et al.* 2015; Lee *et al.* 2016).

The cancer somatic mutation spectrum can be integrated into biomolecular networks (Cheng *et al.* 2014). Cancer evolution may be influenced by somatic mutations in cancer driver genes that cause alterations in other genes. Hofree *et al.* posited a network-based stratification (NBS) method, which applies network propagation to discover cancer subtypes by gathering patients with comparable network mutations (Hofree *et al.* 2013). The hypothesis is that if the mutated genes of 2 tumors are located in similar network regions, they may be very similar. Chuang Liu *et al.* proposed a network embedding-based stratification (NES) approach for identifying clinically relevant patient categories from the somatic mutation spectrum of a large number of patients (Liu *et al.* 2021). Therefore, we can analyze each patient based on their somatic mutation spectrum and the similarities among patients to stratify tumors.

In this work, we introduce a network embedding method for tumor Stratification by integrating Multi-omics data, called NESM. NESM pregroup the samples, integrate the gene features and somatic mutation corresponding to cancer types within each group to construct patient features, and then integrate all groups to obtain comprehensive patient information. The gene features contain network topology information, because it is extracted by integrating DNA methylation, messenger ribonucleic acid (mRNA) expression data and protein–protein interactions (PPIs) through network embedding method. First, we cluster the samples with DNA methylation and mRNA expression data and calculate the Pearson correlation between genes in each cluster. Then, the gene pairs with strong correlation are preserved in PPI. Next, patient features are constructed by integrating corresponding gene features and somatic mutation profiles of cancer types. Finally, a supervised learning method Light Gradient Boosting Machine (lightGBM) is used to classify cancer types based on patient features, while an unsupervised clustering algorithm Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is utilized to divide single cancer subtypes.

Materials and methods

Datasets

The DNA methylation data, mRNA expression data, somatic mutation data, and patient clinical data employed in the study are all downloaded from the TCGA database (Wang *et al.* 2016). We consider 14 cancer types with a total of 5,290 samples (details in Table 1). We collect 5 proven human protein–protein interactomes: (1) by combining 2 publicly available high-quality yeast-two-hybrid (Y2H) datasets, binary PPIs were investigated by high-throughput Y2H systems (Rolland *et al.* 2014; Luck *et al.* 2020); (2) BioPlex V2.016 data on protein complexes discovered by strong affinity purification mass spectrometry techniques (Huttlin *et al.* 2015); (3) Low-throughput or high-throughput experimental tests based on literature from KinomeNetworkX (Cheng *et al.* 2014), Human Protein Resource Database (Peri *et al.* 2004), DbPTM 3.0(Lu *et al.* 2013), PhosphoNetworks (Hu *et al.* 2014), and Phospho.ELM (Dinkel *et al.* 2011); (iv) low-throughput experiments from Signa-Link2.0 are used to create a signaling network (Fazekas *et al.* 2013); and (v) IntAct (Orchard *et al.* 2014), InnateDB (Breuer *et al.* 2013), and low-throughput tests based on literature or protein 3-dimensional structures from BioGRID (Chatr-Aryamontri *et al.* 2015). All genes correspond to Entrez ID and duplicate PPI pairs are removed.

Table 1. Fourteen cancer types and corresponding sample numbers.

Cancer types		Patient number
BLCA	Bladder urothelial carcinoma	406
BRCA	Breast invasive carcinoma	750
CESC	Cervical squamous cell carcinoma	302
COAD	Colon adenocarcinoma	278
HNSC	Head and neck squamous cell carcinoma	504
KIRC	Kidney renal clear cell carcinoma	263
LIHC	Liver hepatocellular carcinoma	369
LUAD	Lung adenocarcinoma	453
LUSC	Lung squamous cell carcinoma	366
READ	Rectum adenocarcinoma	91
SKCM	Stomach adenocarcinoma	468
STAD	Stomach adenocarcinoma	369
THCA	Thyroid carcinoma	498
UCEC	Uterine corpus endometrial carcinoma	173

The overview of NESM is illustrated in Fig. 1. It consists of 3 parts: (1) Samples are clustered using DNA methylation and mRNA expression data. And the Pearson correlation between genes is calculated in each cluster. Then, the gene pairs with strong correlation are preserved in PPI. Next, the network embedding is performed using the struc2vec model. (2) The gene feature generates in step (1) are combined with the somatic mutation spectrum of patients to construct patient features. (3) Patients constructed in step (2) are divided using machine learning methods and validated by survival curves.

Network embedding

The interactions between genes are reflected in the PPI network. To better mine the features of genes in the PPI network, we adopt the struc2vec model (Ribeiro *et al.* 2017) for the vectorization process of PPI network nodes. The Struc2vec model encodes structural similarity by constructing a multilayer graph and generates structural context for nodes. Compared with most algorithms, it can find distant but structurally similar gene pairs, which is more conducive to constructing similar genetic features of patients. The struc2vec model's primary steps are as follows:

Compute structural similarity

The structural similarity $f(x, y)$between each pair of nodes x and y can be denoted as:

$$
\begin{aligned}
f_k(x, y) &= f_{k-1}(x, y) + g(s(R_k(x)), s(R_k(y))), \\
g(s(R_k(x)), s(R_k(y))) &= \frac{\max(s(R_k(x)), s(R_k(y)))}{\min(s(R_k(x)), s(R_k(y)))} - 1,
\end{aligned}
\tag{1}
$$

where $R_k(x)$ represents the set of vertices with $k(k \geq 0)$ distance from the vertex x,$s(R_k(x))$ represents the order sequence of vertex set $R_k(x)$, and $g(s(R_k(x)), s(R_k(y))) > 0$is a function measuring the distance between order sequence $R_k(x)$and $R_k(y)$ and $f_{-1} = 0$.

Construct a hierarchical weighted graph

The edge weights of 2 nodes in the same layer are defined as:

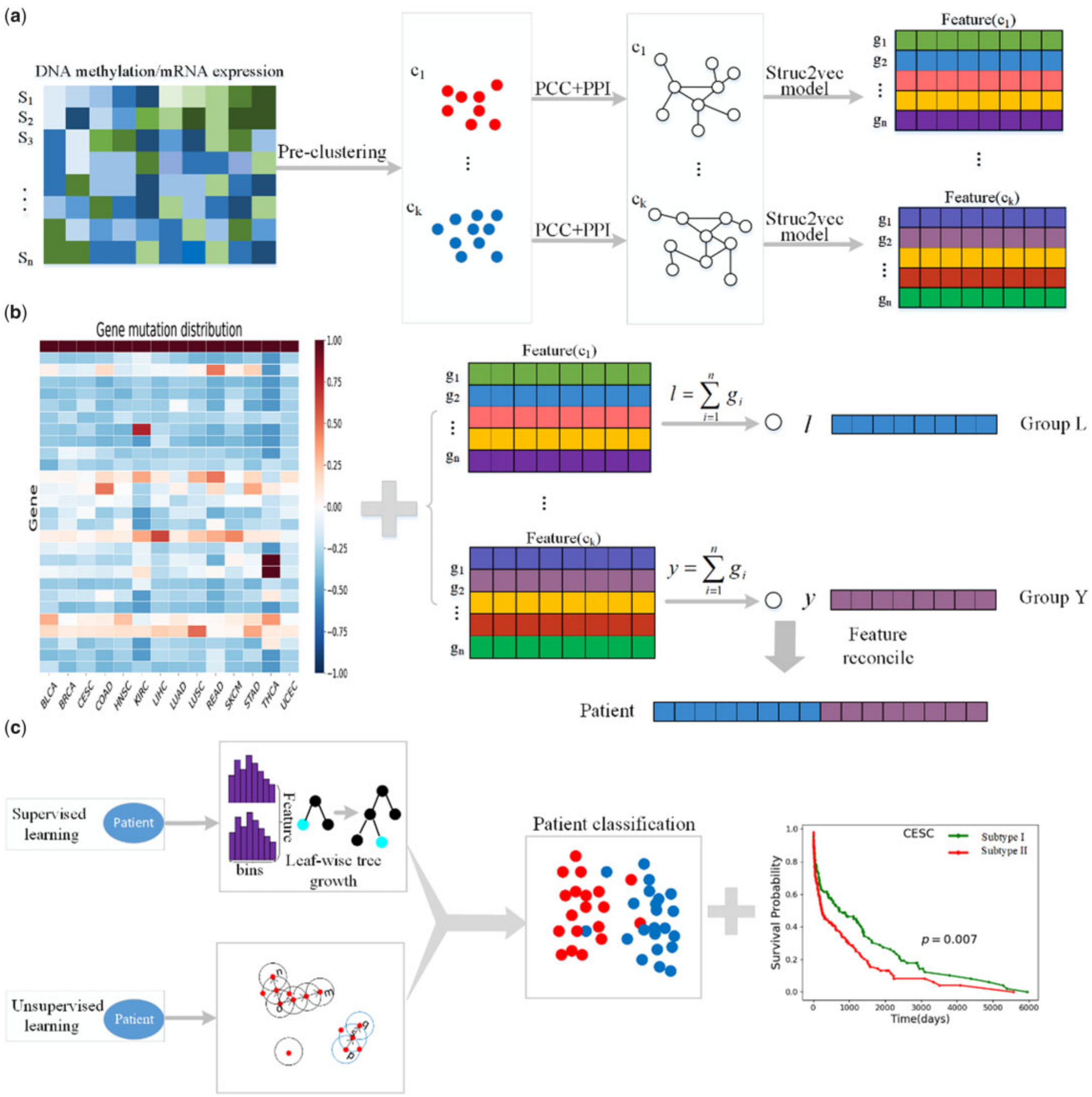

Fig. 1. The overview of NESM.(a) Samples are divided into groups using DNA methylation and mRNA expression data. Pearson correlation is calculated among genes in the group. The gene pairs with strong correlation are then preserved in PPI. Next, the network embedding is performed using the struc2vec model. (b) The gene feature generates in step (a) are combined with the somatic mutation spectrum of the patient to construct the patient features. (c) The patients constructed in step (b) are partitioned using machine learning methods and verified by survival curves.

$$w_k(x,y) = e^{-f_k(x,y)}, k = 0, 1, \ldots k^*, \tag{2}$$

where k^*denotes the original network's diameter.

The same nodes belonging to different levels are connected by directed edges. For a node in the layer k, it will be linked to a node corresponding in the layer $(k-1)$and layer $(k+1)$. The edge weight is defined as:

$$\begin{aligned} w(x_k, x_{k+1}) &= \log(\Gamma_k(x) + e), \\ w(x_k, x_{k-1}) &= 1, \end{aligned} \tag{3}$$

where $\Gamma_k(x)$ is the number of edges at the k layer whose edge weight is greater than the average edge weight of the edge connected to x.

Generate node sequence

We use the biased random walk to carry out random walk in the weighted multilayer graph. It is assumed that the walk takes place in the current layer with the probability of q and jumps to other layers with the probability of $(1-q)$. If it is determined to walk in the current layer, let it be in the layer k, then the probability from node to node is defined as:

$$p_k(x,y) = \frac{e^{-f_k(x,y)}}{Z_k(x)}, \tag{4}$$

where $Z_k(x) = \sum_{y \neq x} e^{-f_k(x,y)}$is the normalized factor x in the kth layer. Through the random walk, each sampled node is more inclined to choose the node with high similarity to the current node structure. If switching to other layers, the probability of selecting $(k-1)$and $(k+1)$is defined as:

$$\begin{aligned} p_k(x_k, x_{k-1}) &= \frac{w(x_k, x_{k-1})}{w(x_k, x_{k-1}) + w(x_k, x_{k+1})} \\ p_k(x_k, x_{k+1}) &= 1 - p_k(x_k, x_{k-1}) \end{aligned} \tag{5}$$

Each random walk sequence in this study has a length of 80 steps. In addition, for each node, create 20 random walk sequences (Fig. 1a). The Skip-Gram model (Mikolov *et al.* 2013) is applied to train node sequences while creating them. A 128-dimensional feature is contained in each gene.

Constructing patient features

To better describe patients, we use mutated genes to construct patient features. We find that the frequency of genes mutations is different in different cancers. As illustrated in Fig. 1b, some genes are mutated in all cancers, while others are mutated only in certain cancers. Therefore, we define a weight to balance the effects of different genetic mutations. The weight is defined as:

$$w(n) = \frac{v_i(n)}{u(n)}, \tag{6}$$

where $u(n)$is the total number of gene $n$mutations in the 14 cancers, and $v_i(n)$ is the total number of gene $n$mutations in cancer v_i. We create a new 128-dimensional feature by fusing mutant genes from the same sample in the same cluster. Finally, the identical samples are spliced across all clusters to create a 1,280-dimensional vector that represents patient features.

Supervised classification and unsupervised clustering models

The constructed patient features are classified using the lightGBM (Ke *et al.* 2017) classification method, which is a supervised approach based on Gradient Boosting Decision Tree (Chen and Guestrin 2016). When classifying cancer types, tumors with the same cancer are taken as positive samples and tumors with other cancers are taken as negative samples. For dichotomies, we use the AUC (Area Under Curve) value as the evaluation index of classification performance. AUC is the area under the Receiver Operating Characteristic (ROC) curve, which is an evaluation index to evaluate the merits of a dichotomous model.

We use DBSCAN clustering to cluster different subtypes of same cancer, which is an unsupervised density clustering method (Ester *et al.* 1996). Given the neighborhood radius δ, the threshold of the number of data objects in the neighborhood $MinPts$, for the cluster M (temporary), the domain of $p(p \in N)$can be calculated using the formula:

$$N_\delta(p) = \{q \in M | d(p,q) \leq \delta\}, \tag{7}$$

where the distance between the node p and $q$is denoted by $d(p,q)$. If $N_\delta(p) \geq MinPts$,p is the central point.

Results

Pan-cancer classification

In this work, we randomly choose 14 cancer types, but we are not limited to 14, from the TCGA database to collect the corresponding clinical information, mRNA expression data, DNA methylation data, and gene mutation data. We preprocess data for each cancer type by obtaining common samples containing all 3 data, normalizing the mRNA expression data and averaging the methylation sites on the same gene. Then, a total of 5,290 samples with gene mutation, DNA methylation, and mRNA expression data in 14 cancer types are obtained (details in Table 1). We generate features by integrating the list of mutated genes and the genetic features obtained through network embedding. Here, each patient is represented by a feature of 1,280 dimensions. To view the distribution status of 14 cancer patients, we visualize patients with 14 cancer types using the t-distributed stochastic neighbor embedding algorithm (Van der Maaten and Hinton 2008) and represented by different colors. We find that most patients of the same type tended to cluster together (Fig. 2). This is due to the fact that different cancer types have different mutation frequencies (Fig. 1b), and patients with the same cancer type are more likely to cluster together, while patients with different cancer types are separated.

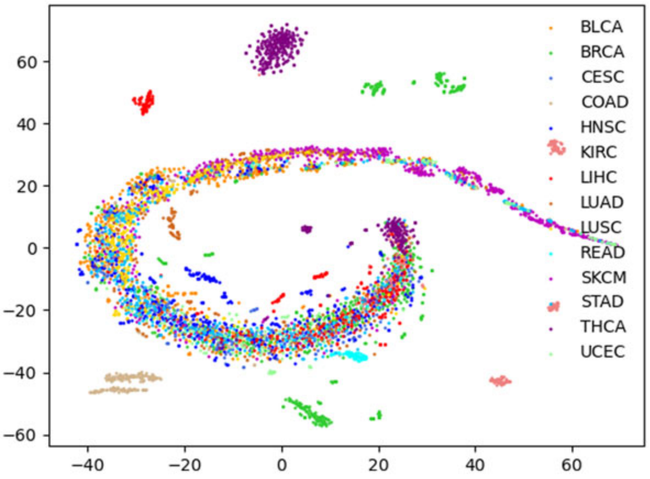

Fig. 2. Visualization of patients using t-SNE.

We use the lightGBM classification algorithm to predict the patient subpopulations and patient features as input of the algorithm to test the feasibility of NESM. When identifying cancer types, the corresponding cancer patients are positive samples, and other cancer patients are negative samples. In the case of colon adenocarcinoma (COAD), patients in COAD cancer are considered as positive samples, while patients from other cancers are negative samples. By using 5-fold cross validation, the positive and negative samples are separated into training and testing sets. The 5-fold cross validations perform 100 times and the average of the results is the final AUC value. Moreover, we compare with 3 latest methods: NES, NBS, and ECC methods. As shown in Fig. 3, our method has obvious advantages in bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), and COAD cancer types. It is slightly lower than the NBS method in cervical squamous cell carcinoma (CESC) and lung adenocarcinoma (LUAD) cancer types and slightly lower than the NES method in uterine corpus endometrial carcinoma (UCEC) cancer. On the whole, our method is better than the other 3 methods. Furthermore, we show that using a single omics data for classification is less effective than using multi-omics fusion (Fig. 4).

Stratification of specific cancer subgroups

Another aim in this work is to classify patients with the same cancer type into the corresponding subtypes. We assemble clinical information on patients with 14 cancer types from the TCGA database and obtain staging information to assess the stratification of

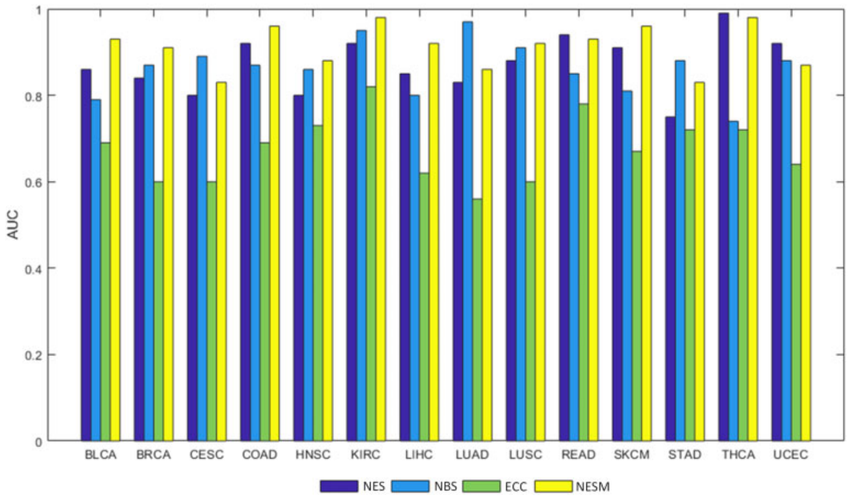

Fig. 3. The AUC values of our NESM method are compared with those of ECC, NBS, and NES methods for 14 cancer types.

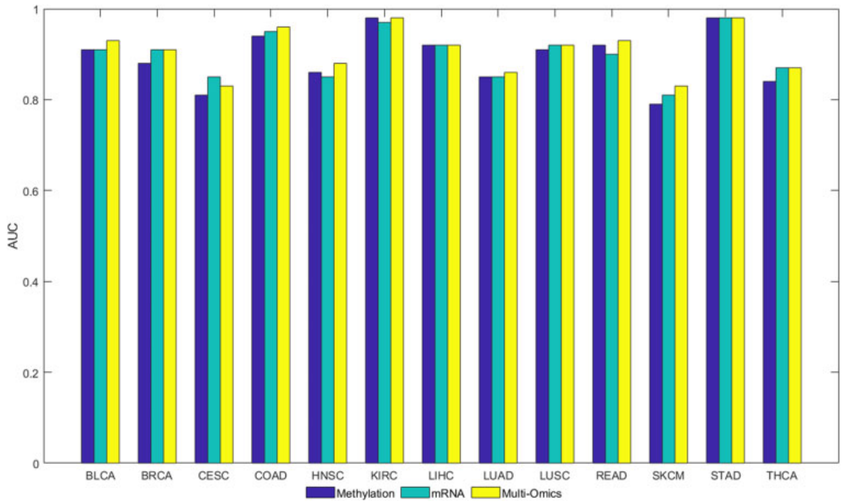

Fig. 4. The AUC values are compared under the NESM framework using mRNA expression, methylation, and multi-omics data.

tumor mutations. The tumor stage refers to the primary tumor and the extent of intraindividual spread. In general, the more extensive the spread, the worse the prognosis. The tumor-node-metastasis (TNM) staging system is the most widely utilized around the world. After T, N, and M are determined in TNM staging, corresponding general staging can be obtained, namely stage I, II, III, and IV. The stage I COAD tumor is taken as an example, and the stage I COAD tumor is labeled as a positive sample. Nonstage I COAD tumor markers are negative samples. We find that the average AUC of most tumor stages is higher than 0.75, while the average AUC of kidney renal clear cell carcinoma (KIRC) tumor stages is 0.66. Figure 5 shows our comparison with the NES method in COAD, BRCA, head and neck squamous cell carcinoma (HNSC), and lung squamous cell carcinoma (LUSC) data. Under COAD data, COAD outperforms NES in all cases. Under BRCA data, NESM is higher than NES at stages I, II, and III. Under HNSC data, NESM is higher than NES at stages I, II, and IV. Under LUSC data, NESM is higher than NES at stages I, III, and IV. In summary, our method has some advantages. (Supplementary Fig. 1 illustrates the staging results of 14 tumors.)

To test the algorithm's robustness, we evaluate the parameters involved in the algorithm. In our method, 2 main parameters affect the algorithm: initial clustering and weight screening. We find that the time cost of the algorithm increases with the increase of parameter value. Therefore, we take α as 3, 4, and 5 and βas 0.4, 0.5, and 0.6, respectively, for discussion. When parameter $\alpha$is 3, 4, and 5, the corresponding AUCs are 0.89, 0.90 and 0.91, respectively. When parameter $\beta$is 0.4, 0.5, and 0.6, the corresponding AUCs are 0.90, 0.91, and 0.90, respectively. The AUC values under different parameters are given in Fig. 6, and the small fluctuation range evidences that NESM is robust.

Based on the above studies, we believe that patients with similar clinical information may be more inclined to cluster together. This means that we can obtain the optimal cluster of patients through an unsupervised learning algorithm, that is, patients are subdivided. We use the DBSCAN clustering algorithm to cluster the same cancer patients. The number of clusters matches the number of subtypes reported in the literature when it comes to identifying medical tumor subtypes. Taking CESC as an example, it is divided into 2 subtypes according to clinical and endocrine features (divided into—type I and type II) or histopathological features (divided into endometrioid, serous or clear cell adenocarcinoma). We generate 2 groups of patients with endometrial carcinoma (CESC) in the data. We also evaluate the clustering

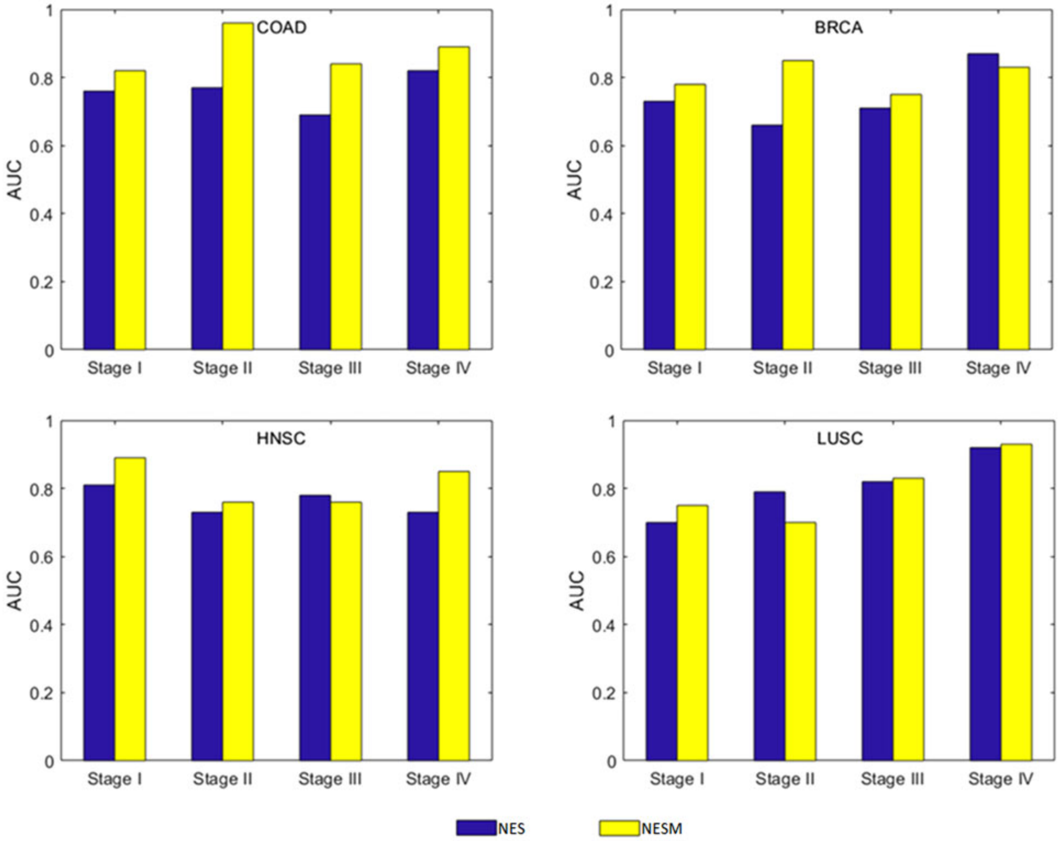

Fig. 5. AUC of patients with NESM and NES tumor stages is compared in COAD, BRCA, HNSC, and LUSC cancer types.

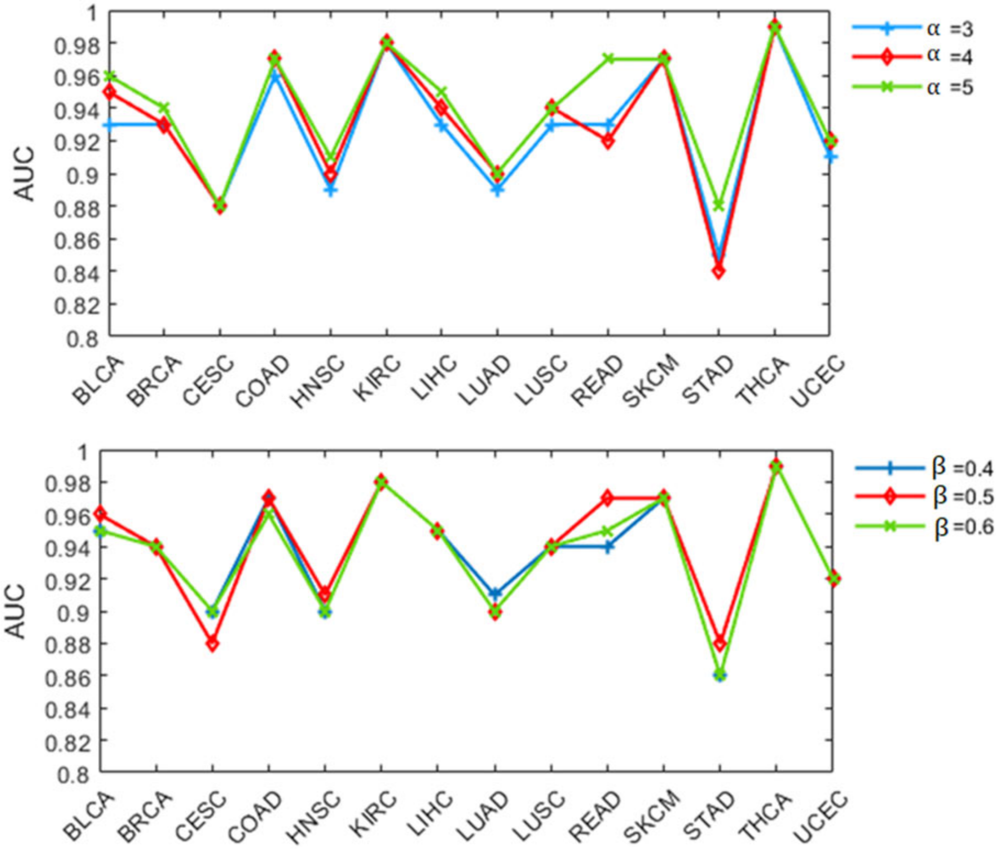

Fig. 6. The line chart shows the effect of NESM on the classification of 14 cancers under different parameters.

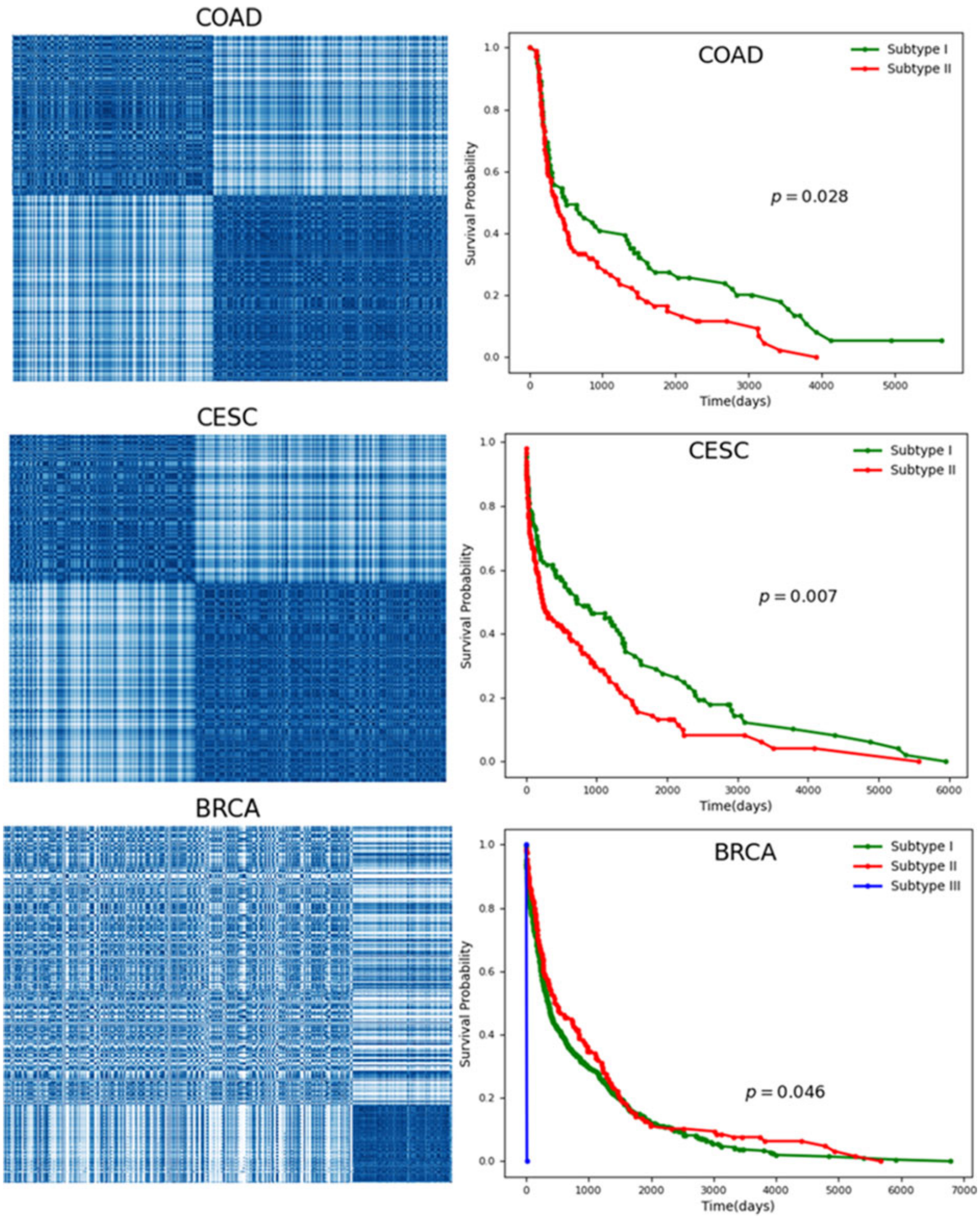

Fig. 7. The results of tumor mutation stratification in COAD, CESC, and BRCA cancers. At the left, patient clusters are presented, with dense patches indicating that comparable individuals should cluster into the same subtype. Survival analysis of patients with different subtypes is shown on the right, and *P*-values are calculated based on the log-rank test.

results by calculating the Silhouette Coefficient and Calinski–Harabasz Index (Supplementary Table 2). For the vast majority of cancer types, patients are closely distributed in the same population. Furthermore, we assess patient survival. As can be seen from the Kaplan–Meier survival diagram in Fig. 7, survival rates of the identified subtypes significantly differ from time and probability. For example, the survival times of the 2 CESC subtypes have noteworthy differences ($P = 7 \times 10^{-3}$). We provide COAD, CESC, and BRCA clustering results in Fig. 7 and carry out survival analysis for each subtype. The lower *P*-values indicate that the subtypes identified by NESM are reliable. In addition, we provide clustering and survival analyses ($P < 0.05$) for other cancers in Supplementary Figs. 2 and 3. Across 14 cancer types, the majority of cancer subtypes identified by NESM are significantly associated with patient survival.

Discussion

Cancer is a multifaceted illness caused by both hereditary and nongenetic components. With the development of technology, multi-omics data has recently been widely used for various cancer types. In this work, NESM constructs patient features by integrating corresponding gene features and somatic mutation profiles of cancer types. Since network topology information is extracted by integrating DNA methylation, mRNA expression data and PPIs through network embedding method, it is contained though the gene features. We apply supervised classification algorithms to classify pan-cancer and individual cancer stages. The experimental results show that the patient features extracted by the NESM method are effective for tumor stratification. When cancer subtypes are subdivided, the vast majority of subtypes identified by the NESM method are significantly associated with patient survival. NESM extracts features mainly from network topology, which is not considered by most methods.

It allows better classification and subdivision of cancers into subtypes than other methods, but it still has some limitations. For example, the choice or construction of a PPI network may have an impact on the NESM model. In addition, the rate of somatic mutation varies greatly among different tumor types. Some tumor types [such as stomach adenocarcinoma (STAD), UCEC, and others] have a high mutation rate, while others have a low mutation rate [such as rectum adenocarcinoma (READ) and BRCA]. In the current NESM framework, we only integrate normal tumor samples that match somatic mutation profiles, DNA methylation, and mRNA expression data. Integration of other types of omics data, including RNA sequencing, individual patient proteomics, and whole-genome sequencing, may further improve the NESM model. Second, the framework of the method is to cluster patients based on patient features extract from specific data sets. This framework can be used to address the tumor stratification problem using a variety of additional algorithms. For example, we can use a graph convolution neural network to improve the prediction accuracy and use other clustering algorithms, including hierarchical clustering and Gaussian mixture model clustering. In future work, it will provide some clues for precision oncology and clinical applications.

Data availability

The code of NESM is available at https://github.com/FengLi12/NESM. Mutation data for PPI and BLCA, BRCA, CESC, COAD, HNSC, KIRC, liver hepatocellular carcinoma, LUAD, LUSC, READ, stomach adenocarcinoma, STAD, thyroid carcinoma, and UCEC are obtained from the literature: doi:10.1093/bioinformatics/btaa1099. DNA methylation and mRNA data are obtained from https://xenabrowser.net/datapages/.

Supplemental material is available at G3 online.

Author contributions

FL and ZS jointly contributed to the design of the study. ZS and YL designed and implemented the method, performed the experiments, and drafted the manuscript. J-XL, JS, and LD participated in the design of the study and performed the statistical analysis. XL and YL contributed to the data analysis. All authors read and approved the final manuscript.

Funding

This work has been supported by the National Natural Science Foundation of China (61902216, 61972236, 61972226, and 61902215) and Natural Science Foundation of Shan-dong Province (No. ZR2018MF013).

Acknowledgments

We are grateful to the anonymous reviewers, whose suggestions and comments contributed to the significant improvement of this article.

Conflicts of interest

No potential conflict of interest was disclosed in this study.

Literature cited

Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ, *et al.* InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. 2013;41(Database issue):D1228–D1233. https://doi.org/10.1093/nar/gks1147.

Chang K, *et al.* The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–1120. https://doi.org/10.1038/ng.2764.

Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, *et al.* The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43(Database issue):D470–D478. https://doi.org/10.1093/nar/gku1204.

Chen T, Guestrin C. 2016. Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 785–794. https://doi.org/10.1145/2939672.2939785.

Cheng F, Jia P, Wang Q, Lin C-C, Li W-H, Zhao Z. Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. Mol Biol Evol. 2014;31(8): 2156–2169. https://doi.org/10.1093/molbev/msu167.

Cheng F, Jia P, Wang Q, Zhao Z. Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. Oncotarget. 2014;5(11):3697–3710. https://doi.org/10.18632/oncotarget.1984.

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F. Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res. 2011;39(Database issue):D261–D267. https://doi.org/10.1093/nar/gkq1104.

Ester M, *et al.* 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD Proceedings. 226–231.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118. https://doi.org/10.1038/nature21056.

Fazekas D, Koltai M, Türei D, Módos D, Pálfy M, Dúl Z, Zsákai L, Szalay-Bekő M, Lenti K, Farkas IJ, *et al.* SignaLink 2—a signaling pathway resource with multi-layered regulatory networks. BMC Syst Biol. 2013;7(1):7–15. https://doi.org/10.1186/1752-0509-7-7.

Ge S-G, Xia J, Sha W, Zheng C-H. Cancer subtype discovery based on integrative model of multigenomic data. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(5):1115–1121. https://doi.org/10.1109/TCBB.2016.2621769.

Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10(11): 1108–1115. https://doi.org/10.1038/nmeth.2651.

Horn H, Lawrence MS, Chouinard CR, Shrestha Y, Hu JX, Worstell E, Shea E, Ilic N, Kim E, Kamburov A, *et al.* NetSig: network-based discovery from cancer genomes. Nat Methods. 2018;15(1):61–66. https://doi.org/10.1038/nmeth.4514.

Hu J, Rho H-S, Newman RH, Zhang J, Zhu H, Qian J. PhosphoNetworks: a database for human phosphorylation networks. Bioinformatics. 2014;30(1):141–142. https://doi.org/10.1093/bioinformatics/btt627.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, *et al.* The BioPlex network: a systematic exploration of the human interactome. Cell. 2015;162(2): 425–440. https://doi.org/10.1016/j.cell.2015.06.043.

Jennings JL, Hudson TJ. International Cancer Genome Consortium (ICGC). Cancer Research. 2016;76(14_Supplement):130–130. https://doi.org/10.1158/1538–7445.AM2016-130.

Ke G, *et al.* Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–3154.

Lee J-H, Zhao X-M, Yoon I, Lee JY, Kwon NH, Wang Y-Y, Lee K-M, Lee M-J, Kim J, Moon H-G, *et al.* Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. Cell Discov. 2016;2(1):1–14. https://doi.org/10.1038/celldisc.2016.25.

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106–114. https://doi.org/10.1038/ng.3168.

Liang C, Shang M, Luo J. Cancer subtype identification by consensus guided graph autoencoders. Bioinformatics. 2021;37(24):4779–4786. https://doi.org/10.1093/bioinformatics/btab535.

Liu C, Han Z, Zhang Z-K, Nussinov R, Cheng F. A network-based deep learning methodology for stratification of tumor mutations. Bioinformatics. 2021;37(1):82–88. https://doi.org/10.1093/bioinformatics/btaa1099.

Liu C, Zhao J, Lu W, Dai Y, Hockings J, Zhou Y, Nussinov R, Eng C, Cheng F. Individualized genetic network analysis reveals new therapeutic vulnerabilities in 6,700 cancer genomes. PLoS Comput Biol. 2020;16(2):e1007701. https://doi.org/10.1371/journal.pcbi.1007701.

Liu H, Zhao R, Fang H, Cheng F, Fu Y, Liu Y-Y. Entropy-based consensus clustering for patient stratification. Bioinformatics. 2017;33(17):2691–2698. https://doi.org/10.1093/bioinformatics/btx167.

Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7(1):523–542. https://doi.org/10.1214/12-AOAS597.

Lu C-T, Huang K-Y, Su M-G, Lee T-Y, Bretaña NA, Chang W-C, Chen Y-J, Chen Y-J, Huang H-D. DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res. 2013;41(Database issue):D295–D305. https://doi.org/10.1093/nar/gks1229.

Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, *et al.* A reference map of the human binary protein interactome. Nature. 2020;580(7803):402–408. https://doi.org/10.1038/s41586-020–2188-x.

Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkänen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology. 1999;57(4):281–286. https://doi.org/10.1159/000012061.

Ma'ayan A. Introduction to network analysis in systems biology. Sci Signal. 2011;4(190):tr5–tr5. https://doi.org/10.1126/scisignal.2001965.

Mikolov T, *et al.* Efficient estimation of word representations in vector space.:1301.3781, https://doi.org/10.48550/arXiv.1301.3781, 2013.

Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci. 2013;110(11):4245–4250. https://doi.org/10.1073/pnas.1208949110.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(Database issue):D358–D363. https://doi.org/10.1093/nar/gkt1115.

Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TKB, Chandrika KN, Deshpande N, Suresh S, *et al.* Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004;32(Database issue):D497–D501. https://doi.org/10.1093/nar/gkh070.

Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet. 2011;378(9805):1812–1823. https://doi.org/10.1016/S0140-6736(11)61539-0.

Ribeiro LF, Saverese PH, Figueiredo DR. 2017. struc2vec: learning node representations from structural identity. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 385–394. https://doi.org/10.1145/3097983.3098061.

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, *et al.* A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–1226. https://doi.org/10.1016/j.cell.2014.10.050.

Ruan P, Wang Y, Shen R, Wang S. Using association signal annotations to boost similarity network fusion. Bioinformatics. 2019;35(19):3718–3726. https://doi.org/10.1093/bioinformatics/btz124.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–2912. https://doi.org/10.1093/bioinformatics/btp543.

Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579–2605.

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–337. https://doi.org/10.1038/nmeth.2810.

Wang Z, Jensen MA, Zenklusen JC. A practical guide to the cancer genome atlas (TCGA). Methods Mol Biol. 2016;1418:111–141. https://doi.org/10.1007/978-1-4939-3578-9_6.

Zhang X, Wang T. Elastic and reliable bandwidth reservation based on distributed traffic monitoring and control. IEEE Trans Parallel Distrib Syst. 2022;33(12):4563–4580. https://doi.org/10.1109/TPDS.2022.3196840.

Zhao L, Yan H. MCNF: a novel method for cancer subtyping by integrating multi-omics and clinical data. IEEE ACM Trans Comput Biol Bioinform. 2020;17(5):1682–1690. https://doi.org/10.1109/TCBB.2019.2910515.

Zhao X-M, Liu K-Q, Zhu G, He F, Duval B, Richer J-M, Huang D-S, Jiang C-J, Hao J-K, Chen L, *et al.* Identifying cancer-related microRNAs based on gene expression data. Bioinformatics. 2015;31(8):1226–1234. https://doi.org/10.1093/bioinformatics/btu811.

Zhong L, *et al.* A laminar augmented cascading flexible neural forest model for classification of cancer subtypes based on gene expression data. BMC Bioinf. 2021;22(1):1–17. https://doi.org/10.1186/s12859-021–04391-2.

Zhu L, Deng S-P, Huang D-S. A two-stage geometric method for pruning unreliable links in protein-protein networks. IEEE Trans Nanobiosci. 2015;14(5):528–534. https://doi.org/10.1109/TNB.2015.2420754.

Communicating editor: T. Matise