Hindawi
Scientific Programming
Volume 2021, Article ID 2292703, 11 pages
https://doi.org/10.1155/2021/2292703

Research Article

Similarity Network Fusion Based on Random Walk and Relative Entropy for Cancer Subtype Prediction of Multigenomic Data

Jian Liu,[1,2] **Wenfeng Liu**,[3] **Yuhu Cheng**,[1,2] **Shuguang Ge**,[1,2] **and Xuesong Wang**[1,2]

[1]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[2]*Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China*
[3]*Department of Information Center, Weihai Ocean Vocational College, Rongcheng 264300, China*

Correspondence should be addressed to Xuesong Wang; wangxuesongcumt@163.com

Received 6 May 2021; Revised 25 July 2021; Accepted 6 August 2021; Published 19 August 2021

Academic Editor: Liang Zhao

It is a crucial task to design an integrated method to discover cancer subtypes and understand the heterogeneity of cancer based on multiple genomic data. In recent years, some clustering algorithms have been proposed and applied to cancer subtype prediction. Among them, similarity network fusion (SNF) can integrate multiple types of genomic data to identify cancer subtypes, which improves the understanding of tumorigenesis. SNF uses a dense similarity matrix to obtain the global information of the data, and the interconnection of samples between different categories will cause noise interference. Therefore, how to construct a more robust dense similarity matrix is an important research content to improve the performance of cancer subtype identification. In this paper, we proposed similarity network fusion based on random walk and relative entropy (R^2SNF) for cancer subtype prediction. Firstly, the random walk algorithm was used to capture the complex relationship between samples in each genomic data. And the transition probability distribution of samples in the network was obtained. If two samples belong to the same class, the transition probability between the two samples is great. On the contrary, if the two samples do not belong to the same class, the transition probability between the two samples is small. In this way, the degree of correlation between samples can be well obtained, thereby reducing the noise interference caused by the interconnection of samples between different categories. Secondly, relative entropy was used to calculate the difference in the transition probability distribution between samples to construct a better dense similarity matrix which contains structural similarity information between samples. Thirdly, we iteratively fused the obtained dense similarity matrix with the KNN similarity matrix to construct the fused similarity matrix of all genomic data. Finally, by using spectral clustering, the fused similarity matrix was grouped into multiple clusters, which indicates the cancer subtypes. Experiments on seven cancer omics datasets show that the R^2SNF algorithm performs well in identifying cancer subtypes.

1. Introduction

With the rapid development of high-throughput technology, a large amount of genomic data has been generated, including gene expression data, DNA methylation data, and DNA copy number variation data. In particular, The Cancer Genome Atlas (TCGA) [1] database researches different genome, transcriptome, and epigenome information of more than 1,100 patients from more than 34 cancer types. These data have brought unprecedented opportunities to cancer research, such as driven gene selection [2] and cancer subtype prediction, so that cancer can be controlled more thoroughly and comprehensively.

Various types of genomic data are closely related to the occurrence and development of cancer. In general, cell growth and differentiation are regulated by the gene expression level, and the changes in the gene expression level will lead to transformation from normal cells to cancer cells [3]. DNA single-nucleotide polymorphisms and copy number variations in the genome affect gene instability and cancer gene activation through gene amplification or cancer suppression [4]. DNA methylation in epigenetic variation is

also common in cancer genomes. Genome-wide hypomethylation can lead to genome instability. The hypomethylation of CpG islands is also related to the inactivation of cancer suppressor genes [5]. At present, many studies have attempted to use these genomic data to predict cancer subtypes. However, the cancer genome is regulated by a variety of molecular mechanisms, the complexity and independence of which make it difficult to discover the relationship between the cancer genome and the cancer phenotype. Therefore, integrating different genomic data to capture the complexity of phenotypes and the heterogeneity of biological processes [6, 7] is the current trend in predicting cancer subtypes.

In the past few decades, many genomic data integration algorithms have been extensively developed. For example, Shen et al. [8] proposed a joint latent variable model named as iCluster, which combines the correlation between different types of genomic data and the variance-covariance structure within the data type to mine potential cancer subtypes. Akavia et al. [9] proposed an algorithm based on the Bayesian network to integrate the matching chromosome copy number and gene expression data of tumor samples to identify driving mutations and their influence processes. Liang et al. [10] proposed a multimodal deep belief network algorithm, which encodes the relationship between features of each genomic data as a multilayer network of hidden variables and then fuses common features to cluster cancer into different subtypes. Speicher and Pfeifer [11] added regularization constraints in the optimization process of multikernel learning to avoid overfitting and used one kernel for each genome data type to solve the problem of kernel function and parameter selection. Wang et al. [12] proposed a multiplexed network, which integrates heterogeneous genomic data by using the links between each node in a network slice and its corresponding nodes in each other network slice. Van et al. [13] used sequencing matrix decomposition to represent genomic data and identify cancer subtypes based on mutations and gene expression characteristics. Zhang and Ma [14] proposed a regularized multiview subspace clustering method to integrate gene expression data with the protein interaction network of dynamic modules. Network-based stratification (NBS) [15, 16] method combines genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes. And the gene interaction network is constructed by protein-protein interactions (PPI). Simultaneous rank matrix factorization (SRF) [13] method approaches the subtyping problem by decomposing patient-mutation and patient-expression data into ranked factors.

Among these integrated algorithms, Wang et al. [6] proposed a very effective cancer subtype identification algorithm—similarity network fusion (SNF). SNF consists of three stages: network construction, network fusion, and clustering. In the network construction stage, the Euclidean distance of each omics data is used to construct a patient similarity network. In the network fusion stage, the information dissemination theory is used to perform nonlinear iterative fusion of the constructed network. Finally, the spectral clustering algorithm is used for clustering. SNF integrates mRNA expression data, DNA methylation data, and miRNA expression data and establishes a cancer subtype prediction model on five cancer datasets.

At present, many studies have improved and expanded SNF. Xu et al. [17] proposed a weighted similarity network fusion algorithm, which uses a complex miRNA-TF-mRNA regulatory network to identify cancer subtypes. In order to solve the problem that SNF is only applicable to data types containing continuous values, Yang et al. [18] used the random walk method to smooth the discrete somatic mutation data and incorporated the smoothed data into the SNF algorithm so that SNF can fuse discrete data. Yang et al. [19] proposed a deep subspace fusion clustering algorithm, which used the methods of self-encoding and data self-expression to guide the deep subspace model, which can effectively express the discriminant similarity between samples, thereby realizing the difference transfer between clustering clusters and the enhancement of compactness within clustering clusters. In view of the superior performance of SNF, it has become one of the most popular algorithms for cancer subtype identification. Therefore, this paper improves SNF from the perspective of similarity matrix construction, aiming to further improve the recognition effect of SNF on cancer subtypes.

After SNF completes network fusion, it needs to be clustered through spectral clustering [20]. The essence of spectral clustering is to map the Laplacian matrix so that the samples in the original space that are not easy to handle can be easily processed in the mapped space. The Laplacian matrix is calculated by the similarity matrix, so the construction of the similarity matrix is the key to SNF. SNF constructs two similarity matrices for each genomic data, dense similarity matrix and sparse similarity matrix, which are used to capture global and local information of genomic data, respectively. In SNF, K-nearest neighbor (KNN) algorithm is used to construct the sparse similarity matrix. KNN algorithm is the most commonly used and effective sparse similarity matrix construction method. All samples in the dense similarity matrix have connecting edges. In spectral clustering, the interconnection of samples of different categories in the dense similarity matrix will cause noise interference and affect the segmentation effect of spectral clustering. Therefore, how to optimize the dense similarity matrix has become a major problem faced by SNF.

In this paper, we proposed similarity network fusion based on random walk [17] and relative entropy (R^2SNF) for cancer subtype prediction. Random walk and relative entropy are used to measure the similarity between samples to construct a more robust dense similarity matrix on each genomic data. The similarity matrix construction method based on random walk measures the transition probability of a sample walking along a randomly selected adjacent edge to reach other samples, thereby forming a transition probability distribution of this sample. In order to better measure the similarity between samples, the relative entropy is used to calculate the difference of the transition probability distribution of them, and the similarity between them is obtained: the greater the difference between two probability

distributions is, the less similar the corresponding samples are, and vice versa. The dense similarity matrix construction method is to establish a random walk point on the basis of the conventional dense similarity matrix. It uses the difference in the transition probability distribution between samples to measure the similarity of two samples so that similar samples have a larger similarity value, and samples that are not of the same class have a smaller similarity value. Thus, a more robust dense similarity matrix is obtained. In our R^{2}SNF, we use the dense similarity matrix obtained above and the sparse similarity matrix obtained by the KNN algorithm to perform similarity network fusion for different genomic data. Finally, we use spectral clustering to cluster the fusion similarity matrix. Experimental results on multiple genomic data show that R^{2}SNF can identify biologically significant cancer subtypes.

2. Methods

In this section, we will introduce our algorithm similarity network fusion based on random walk and relative entropy (R^{2}SNF) in detail. Firstly, the probability distribution of random walk from one sample in each genomic data to other samples in the network is calculated. Secondly, relative entropy is used to calculate the difference of the probability distribution of the two samples, and the robust dense similarity matrix is constructed. Thirdly, similarity network fusion between the constructed robust dense similarity matrix and the KNN similarity matrix is performed to obtain the fused similarity matrix. Finally, spectral clustering is used for clustering the fused similarity matrix.

2.1. Construction of the Random Walk Model. Random walk [21] is a random process model that can simulate the interaction between samples in the network. The random walk on the graph can be regarded as a Markov chain of randomly selected nodes. After years of development, a variety of random walk algorithms have been produced. Here, we use the random walk with restart (RWR) algorithm proposed by Tong et al. [22].

Given a set of cancer genomic data $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^v, \ldots, \mathbf{X}^V\}$, $\mathbf{X}^v \in \mathbb{R}^{n \times m^v}$, where V represents the number of genomic data, $\mathbf{X}^v$ is the vth genomic data in $\mathbf{X}$, m^v represents that the vth genomic data have m features, and n represents the number of samples. For each genomic data $\mathbf{X}^v$, starting from the ith sample x_i^v, each step of the RWR faces two choices: choose the adjacent sample with the probability of α or return to the starting sample with the probability of $1 - \alpha$; then, the sample x_i^v will transfer to any sample and reach a stable state at the time $t + 1$. According to the Markov decision process, the current state of the system is only related to the state at the previous moment. Therefore, the stable state vector $\mathbf{r}_{t+1}^v(\mathbf{x}_i^v)$ at the time $t + 1$ can be defined as

$$\mathbf{r}_{t+1}^v(\mathbf{x}_i^v) = \alpha \mathbf{r}_t^v(\mathbf{x}_i^v)\mathbf{A}^v + (1-\alpha)\mathbf{r}_0^v(\mathbf{x}_i^v), \tag{1}$$

where $\mathbf{r}_t^v(\mathbf{x}_i^v)$ represents the state vector at the time t, $\mathbf{r}_0^v(\mathbf{x}_i^v)$ is the initialization vector with the ith element being 1 and the remaining elements being 0, and $\mathbf{A}^v \in \mathbb{R}^{n \times n}$ represents the transition probability matrix of each genomic data. Under normal circumstances, the probability transition matrix of the random walk on the graph can be represented by the adjacency matrix after data normalization. We adopt the following ideas to construct the transition probability matrix $\mathbf{A}^v$.

Firstly, we construct the similarity matrix $\mathbf{W}^v \in \mathbb{R}^{n \times n}$ for each genomic data by

$$\mathbf{W}^v(i,j) = \exp\left(-\frac{\rho^2\left(\mathbf{x}_i^v, \mathbf{x}_j^v\right)}{\mu \varepsilon_{i,j}}\right), \tag{2}$$

where $\mathbf{W}^v(i,j)$ represents the similarity between sample $\mathbf{x}_i^v$ and sample $\mathbf{x}_j^v$, μ is an empirical hyperparameter, and $\rho^2(\mathbf{x}_i^v, \mathbf{x}_j^v)$ is the Euclidean distance between samples $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$. $\varepsilon_{i,j}$ can be defined as

$$\varepsilon_{i,j} = \frac{1}{3}\left(\text{mean}\left(\rho\left(\mathbf{x}_i^v, \mathbf{N}_i^v\right)\right) + \text{mean}\left(\rho\left(\mathbf{x}_j^v, \mathbf{N}_j^v\right)\right) + \rho\left(\mathbf{x}_i^v, \mathbf{x}_j^v\right)\right), \tag{3}$$

where $\text{mean}(\rho(\mathbf{x}_i^v, \mathbf{N}_i^v))$ denotes the average of the distances between the sample $\mathbf{x}_i^v$ and its neighbors.

In the process of random walk, $\mathbf{A}^v$ is a probability transition matrix, which needs to meet the condition $\sum_j \mathbf{A}^v(i,j) = 1$. We can get $\mathbf{A}^v$ by normalizing $\mathbf{W}^v$:

$$\mathbf{A}^v = \left(\mathbf{D}^v\right)^{-1}\mathbf{W}^v, \tag{4}$$

where $\mathbf{D}^v$ is the degree matrix, and its diagonal elements satisfy $\mathbf{D}^v(i,j) = \sum_j \mathbf{W}^v(i,j)$.

2.2. Construction of the Similarity Matrix Based on Relative Entropy. After calculating the stable state transition probability distribution $\mathbf{r}^v$ from the RWR in Section 2.1, the similarity $\mathbf{S}^v(\mathbf{x}_i^v, \mathbf{x}_j^v)$ between the sample $\mathbf{x}_i^v$ and sample $\mathbf{x}_j^v$ is usually defined as [23]

$$\mathbf{S}^v\left(\mathbf{x}_i^v, \mathbf{x}_j^v\right) = \mathbf{r}^v_{\mathbf{x}_i^v, \mathbf{x}_j^v} + \mathbf{r}^v_{\mathbf{x}_j^v, \mathbf{x}_i^v}, \tag{5}$$

where $\mathbf{r}^v_{\mathbf{x}_i^v, \mathbf{x}_j^v}$ is the probability of starting from $\mathbf{x}_i^v$ and arriving at $\mathbf{x}_j^v$ via random walk. However, this method only considers the probability value of the random walk between the two samples and ignores the structural similarity between them.

In order to better measure the similarity between samples, the difference in the transition probability distribution of two nodes is used to define the structural similarity. We use the relative entropy to construct the dense similarity matrix [24]. Relative entropy, also known as Kullback–Leibler (KL) divergence [25], is a method to describe the difference between two probability distributions. Here, relative entropy is used to calculate the difference of the transfer probability distribution of different samples.

For sample $\mathbf{x}_i^v$, the transition probability distribution $\mathbf{r}^v(\mathbf{x}_i^v)$ of reaching any other sample to reach a stable state after random walk can be written as

$$\mathbf{r}^v\left(\mathbf{x}_i^v\right) = \left[\mathbf{r}^v\left(\mathbf{x}_i^v, \mathbf{x}_1^v\right), \mathbf{r}^v\left(\mathbf{x}_i^v, \mathbf{x}_2^v\right), \ldots, \mathbf{r}^v\left(\mathbf{x}_i^v, \mathbf{x}_n^v\right)\right], \tag{6}$$

where n is the number of samples and $\mathbf{r}^v(\mathbf{x}_i^v, \mathbf{x}_j^v)$ is the new probability of starting from $\mathbf{x}_i^v$ and arriving at $\mathbf{x}_j^v$ via random walk. $\mathbf{r}^v(\mathbf{x}_i^v, \mathbf{x}_j^v)$ can be defined as

$$\mathbf{r}^v\left(\mathbf{x}_i^v, \mathbf{x}_j^v\right) = \frac{\mathbf{r}^v_{\mathbf{x}_i^v, \mathbf{x}_j^v}}{\sum_{k=1}^n \mathbf{r}^v_{\mathbf{x}_i^v, \mathbf{x}_k^v}}. \tag{7}$$

For the transition probability distribution $\mathbf{r}^v(\mathbf{x}_i^v)$ and $\mathbf{r}^v(\mathbf{x}_j^v)$ of any two samples $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$, respectively, the relative entropy can be defined as

$$D_{\mathrm{KL}}\left(\mathbf{r}^v(\mathbf{x}_i^v)\middle\|\mathbf{r}^v\left(\mathbf{x}_j^v\right)\right) = \sum_k^n \mathbf{r}^v(\mathbf{x}_i^v, \mathbf{x}_k^v)\log_2\frac{\mathbf{r}^v(\mathbf{x}_i^v, \mathbf{x}_k^v)}{\mathbf{r}^v\left(\mathbf{x}_j^v, \mathbf{x}_k^v\right)}. \tag{8}$$

When $a = 0$ or $b = 0$, we define $\log_2(a/b) = 0$.

Relative entropy is an asymmetric measure; that is, $D_{\mathrm{KL}}(\mathbf{r}^v(\mathbf{x}_i^v)\|\mathbf{r}^v(\mathbf{x}_j^v)) \neq D_{\mathrm{KL}}(\mathbf{r}^v(\mathbf{x}_j^v)\|\mathbf{r}^v(\mathbf{x}_i^v))$. Therefore, the probability distribution difference matrix is defined as $\mathbf{C}^v$; then, the difference between any two probability distributions is $\mathbf{C}^v(i, j)$:

$$\mathbf{C}^v(i, j) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(\mathbf{r}^v(\mathbf{x}_i^v)\middle\|\mathbf{r}^v\left(\mathbf{x}_j^v\right)\right) + D_{\mathrm{KL}}\left(\mathbf{r}^v\left(\mathbf{x}_j^v\right)\middle\|\mathbf{r}^v(\mathbf{x}_i^v)\right)\right). \tag{9}$$

Finally, $\mathbf{C}^v$ is transformed into a similarity matrix $\mathbf{S}^v$, where the elements are defined as $\mathbf{S}^v(i, j)$:

$$\mathbf{S}^v(i, j) = 1 - \frac{\mathbf{C}^v(i, j)}{\mathbf{C}^v_{\max}}, \tag{10}$$

where $\mathbf{C}^v_{\max}$ is the maximum in $\mathbf{C}^v$. From equation (8), we can get the following: when the transition probability distribution between samples $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$ differs greatly, that is, the value of $\mathbf{C}^v(i, j)$ is very large, a smaller value of $\mathbf{S}^v(i, j)$ is assigned. On the contrary, when the difference of the transition probability distribution between samples $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$ is small, that is, the value of $\mathbf{C}^v(i, j)$ is small, a great value of $\mathbf{S}^v(i, j)$ is assigned. Thus, the construction of the similarity matrix based on relative entropy is realized.

2.3. Similarity Network Fusion Based on Random Walk and Relative Entropy (R^2SNF). Through the above two steps, the similarity matrix $\mathbf{S}^v$ is obtained. In the similarity network fusion stage, we use $\mathbf{S}^v$ as a dense similarity matrix to obtain the global structure between samples and use the KNN similarity matrix to capture the local structure.

For any samples $\mathbf{x}_i^v$, KNN defines the similarity matrix $\mathbf{K}^v$ between $\mathbf{x}_i^v$ and its k most similar samples. The element $\mathbf{K}^v(i, j)$ in $\mathbf{K}^v$ is defined as

$$\mathbf{K}^v(i, j) = \begin{cases} \dfrac{\mathbf{W}^v(i, j)}{\sum_{k\in\mathbf{N}_i^v}\mathbf{W}^v(i, k)}, & j \in \mathbf{N}_i^v, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where $\mathbf{N}_i^v$ is the neighbors of $\mathbf{x}_i^v$.

Assume that there is a total of V genomic data to be integrated. In the same way as SNF, we performed nonlinear iterative fusion for dense similarity matrix $\mathbf{S}^v$ and sparse similarity matrix $\mathbf{K}^v$ of each dataset. The fusion process can be described as

$$\widetilde{\mathbf{S}}^v = \mathbf{K}^v \times \left(\frac{\sum_{k\neq v}\mathbf{S}^k}{V-1}\right) \times \left(\mathbf{K}^v\right)^{\mathrm{T}}, \quad v = 1, 2, \ldots, V. \tag{12}$$

According to equation (12), we can obtain the similarity matrix $\widetilde{\mathbf{S}}^v$ of the cross-diffusion of the vth genomic data with other data. Then, the final fused similarity matrix $\mathbf{S}$ can be obtained by averaging all $\widetilde{\mathbf{S}}^v$:

$$\mathbf{S} = \frac{1}{V}\sum_{v=1}^{V}\widetilde{\mathbf{S}}^v. \tag{13}$$

2.4. Spectral Clustering on the Fused Similarity Matrix. Suppose we want to identify c cancer subtypes from multiple genomic data, so we need to use spectral clustering to cluster cancer samples into c clusters. For the ith sample, we defined a cluster indicator vector $\mathbf{y}_i \in \{0, 1\}$. When the ith sample belongs to the jth cluster, $\mathbf{y}_i(j) = 1$; otherwise, $\mathbf{y}_i(j) = 0$. The cluster indicator matrix can be written as $\mathbf{Y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)$.

With the fused similarity matrix $\mathbf{S}$, spectral clustering can be performed by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{U}} \mathrm{tr}\left(\mathbf{U}^T\mathbf{L}\mathbf{U}\right) \\ &\text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}, \end{aligned} \tag{14}$$

where $\mathbf{U} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$, $\mathbf{U} \in \mathbb{R}^{n\times c}$, is the scaled partition matrix. According to the fused similarity matrix $\mathbf{S}$, $\mathbf{L}$ as the normalized Laplacian matrix can be defined as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the degree matrix, which satisfies $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, $d_i = \sum_{j=1}^n \mathbf{S}(i, j)$. In this way, we can capture the global structure of the fused similarity matrix through spectral clustering.

3. Results and Discussion

3.1. Datasets and Survival Analysis. In this paper, we tested the proposed algorithm on three types of genomic data, that is, mRNA expression data, miRNA expression data, and DNA methylation data. The cancer types we tested include glioblastoma multiforme (GBM), breast invasive carcinoma (BIC), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC), and colon adenocarcinoma (COAD). The above data can be downloaded from the TCGA website [5]. In addition, we also conducted experiments on the BREAST cancer and LUNG cancer datasets in [26]. The detailed information of the cancer multigenomic datasets is shown in Table 1.

This paper conducts survival analysis based on the cancer subtypes obtained by clustering to verify the survival differences among samples of different cancer subtypes found by the proposed algorithm. In statistics, hypothesis testing is usually used to quantify whether there are differences between different survival curves. Here, the Cox log-rank test [27] is used to calculate the p value. Cox log-rank test is a nonparametric hypothesis test, which is often used to assess the importance of differences in survival between subtypes.

Table 1: Detailed information on seven types of cancer multigenomic datasets.

Cancer type	Number of genes			Number of samples
	mRNA	Methylation	miRNA	
GBM	12042	1305	534	215
BIC	17814	23094	354	105
KRCCC	17899	24960	329	122
LSCC	12042	23074	352	106
COAD	17814	23088	312	92
BREAST	20531	5000	1046	622
LUNG	20531	5000	1046	337

The p value indicates that the observed difference in survival is the likelihood of an incident occurring by chance. Therefore, the smaller the p value is, the better the experimental effect is. In addition, the Kaplan–Meier estimation method [28] is usually used to estimate the survival function and further obtain the Kaplan–Meier survival curve. The x-axis of the survival curve is the time from the beginning of observation to the last observation time point. The y-axis is the survival rate of the survival sample. The curve represents the development of the event.

3.2. Experimental Results. We compared the proposed algorithm R^2SNF with several cancer subtype prediction methods, e.g., SNF [6], LRAcluster [29], iClusterPlus [30], pattern fusion analysis (PFA) [31], affinity network fusion (ANF) [32], and multiview clustering based on Stiefel manifold (MCSM) [33], to verify its effectiveness. In order to verify whether the relative entropy in the R^2SNF algorithm can improve the prediction results of cancer subtypes, we remove the relative entropy from R^2SNF and use equation (5) to construct the similarity matrix. We name the above algorithm as similarity network fusion based on random walk (RSNF). A brief introduction to these methods is as follows:

(i) SNF first uses the exponential similarity kernel method to define the similarity between the sample points of each genomic data. It uses the KNN method to define a dense similarity matrix and a sparse similarity matrix. Then, the information transfer model is proposed to fuse the above two similarity matrices, and the fused similarity matrix can be obtained by updating iteratively. Finally, spectral clustering is used to cluster the fused similarity matrix.

(ii) LRAcluster is a dimensional reduction and clustering method for multigenomic data based on low-rank approximation. It can deal with a variety of distributed data classes and guarantee the orthogonality of the low-dimensional space. It is suitable for clustering analysis of large-scale multigenomic data and has been widely concerned and applied.

(iii) iClusterPlus considers that different variable types should follow different linear probability relationships. Then, it builds a joint sparse model to complete the task of sample clustering and feature selection.

(iv) PFA uses the local information extraction method to project each genomic data in a low-dimensional space and builds a dynamic collimation method based on the idea of manifold learning. Then, it integrates the low-dimensional spatial information into a feature space containing information from different genomic data. Finally, the K-means method is used to cluster the samples.

(v) ANF first constructs a patient affinity network from each omics data and then fuses all individual networks to obtain a more robust one. In order to make the patient affinity network robust to noise, ANF mainly employs two nonlinear k-nearest-neighbor- (kNN-) based transformations: kNN Gaussian kernel and kNN graph.

(vi) MCSF establishes a binary optimization model for the simultaneous clustering problem. Then, the optimization problem is solved by the linear search algorithm based on the Stiefel manifold. Finally, it integrated the clustering results obtained from multiomics data by using the k-nearest neighbor method.

(vii) RSNF obtains the probability of each sample starting from one sample and arriving at another via random walk, calculates the similarity matrix according to the random walk probability between the two samples, and finally performs similarity network fusion according to SNF.

Since R^2SNF is an improved version of SNF, in order to make a more intuitive comparison and analysis, we used the number of clusters suggested in SNF, that is, GBM is clustered into 3 categories, BIC is clustered into 5 categories, KRCCC is clustered into 3 categories, LSCC is clustered into 4 categories, and COAD is clustered into 3 categories. For the BREAST and LUNG datasets, we also used the cancer subtype determination method in SNF to determine the number of their cancer subtypes as 3 and 2, respectively.

The specific experimental results of R^2SNF and other methods on the seven cancer multigenomic datasets are shown in Table 2. Compared with RSNF, R^2SNF had better results on the other six datasets except for KRCCC data. This shows that using relative entropy to calculate the probability distribution difference between samples is beneficial to the construction of the similarity matrix. Compared with SNF, R^2SNF has smaller p values on all datasets except for COAD. The results of RSNF on GBM, BIC, KRCCC, and LSCC are better than SNF, especially on KRCCC and LSCC data, but slightly worse than SNF on other data, which indicates that only using the probability obtained by random walk between samples to construct the similarity matrix also has a certain effect on cancer subtypes. Compared with other algorithms, R^2SNF has the best results on the whole. Only on BIC data, MCSM algorithm is better than R^2SNF.

Figure 1 shows the Kaplan–Meier survival curve of cancer subtypes identified by R^2SNF on seven cancer

Table 2: Comparison of p values between R^{2}SNF and other algorithms on seven cancer multigenomic datasets.

Cancer type	Methods							
	R^{2}SNF	SNF	LRAcluster	iClusterPlus	PFA	ANF	MCSM	RSNF
GBM	**2.4E − 05**	2.0E − 04	3.5E − 04	3.0E − 03	8.0E − 05	5.8E − 04	1.1E − 03	1.2E − 04
BIC	1.1E − 04	1.1E − 03	4.3E − 02	3.5E − 02	9.3E − 03	3.6E − 04	**3.1E − 05**	1.2E − 04
KRCCC	7.0E − 03	2.9E − 02	3.2E − 02	1.1E − 01	7.5E − 03	2.9E − 02	8.0E − 02	**2.1E − 04**
LSCC	**1.5E − 05**	2.0E − 02	5.7E − 02	5.2E − 02	4.0E − 03	8.9E − 03	1.6E − 02	2.0E − 04
COAD	1.8E − 03	**1.3E − 03**	9.9E − 03	5.0E − 02	6.7E − 02	9.0E − 03	3.6E − 01	6.0E − 03
BREAST	**4.5E − 09**	1.0E − 08	3.0E − 01	1.5E − 01	3.6E − 07	1.9E − 08	7.4E − 07	2.3E − 08
LUNG	**5.6E − 03**	1.1E − 02	4.6E − 01	6.9E − 01	2.0E − 01	1.0E − 02	8.0E − 02	8.2E − 02

The best results have been highlighted in bold.

genomic datasets. It can be seen that, on GBM, KRCCC, LSCC, COAD, BREAST, and LUNG, there is a big difference between the cancer subtypes recognized by R^{2}SNF, indicating that R^{2}SNF is an effective method for identifying cancer subtypes. On BIC data, SNF suggested to divide it into 5 cancer subtypes. As shown in Figure 1(b), R^{2}SNF is not very effective when divided into 5 subtypes, but it can clearly divide it into 3 subtypes. Moreover, the p value of SNF on the BIC data is lower than the p value of SNF. Therefore, we recommend that BIC should be divided into 3 subtypes. The number of clusters given in the BREAST dataset in [26] is 3, which can be found in Figure 1(f). This further verifies our conclusion.

3.3. Analysis on the GBM Dataset. Glioblastoma multiforme (GBM) is the most common and lethal malignant primary brain tumor in adults and is one of a group of tumors known as gliomas. Many studies have carried out research on GBM at the molecular level. And clinically, some studies have given definite cancer subtypes and corresponding treatment plans. For example, based on mRNA expression data, Verhaak et al. [34] divided GBM into four cancer subtypes: mesenchymal, classical, neural, and proneural. In [35], according to the difference of the CpG island methylator phenotype (CLMP), GBM was divided into two cancer subtypes: G-CLMP and non-G-CLMP.

On GBM data, we counted the distribution of clustering results obtained by R^{2}SNF on the cancer subtypes determined in the above two studies and summarized the results in Table 3. Table 3 shows that the patients in subtype 1 are more than in subtype 3. Most patients in subtype 1 are grouped into non-G-CLMP (accounted for 99.3%); also, they are distributed on four subtypes in [34]. Subtypes 2 and 1 have similar distributions. It is worth noting that most of the 19 patients with subtype 3 are of the G-CLMP subtype (accounted for 73.7%), and all of them are of the proneural subtype.

To further analyze the obtained cancer subtypes by R^{2}SNF, the clinical data for all patients of GBM were downloaded from the cBio Cancer Genomics Portal database. We drew a boxplot of the age distribution of patients in the three cancer subtypes (Figure 2). Figure 2 proves that the cancer subtypes identified by R^{2}SNF have a clear age distribution difference. Combining Figures 1 and 2, we can find that the age of patients in subtype 3 with the best survival advantage in Figure 1 is also lower than that of patients in subtypes 1 and 2.

Furthermore, we drew Kaplan–Meier survival curves of GBM patients' response to the drug temozolomide (TMZ) in Figure 3. The patients within the three cancer subtypes were divided into two parts: patients treated with drug TMZ and those not treated with drug TMZ. TMZ is a drug that is commonly used to treat GBM, but only responds well to a subset of patients. The p values of survival analysis in the Cox log-rank model of the three cancer subtypes are 5.42×10^{-6}, 3.78×10^{-4}, and 0.36, respectively, which indicate that TMZ has no effect on the patients in cancer subtype 3.

In summary, subtype 3 of GBM identified by R^{2}SNF has the following characteristics. First, most of the patients with subtype 3 are of the G-CLMP subtype, and all of them are of the proneural subtype. Second, the age of patients in subtype 3 with the best survival advantage is also lower than that of patients in subtypes 1 and 2. Third, TMZ has no effect on the patients in cancer subtype 3. Therefore, we believe that subtype 3 identified by R^{2}SNF is a biologically significant cancer subtype. In addition, it can be inferred that we get a potential cancer subtype, which contains patients belonging to both G-CLAMP and Proneural. This verified the study reported by Brennan et al. that the proneural subtype granted by the G-CIMP phenotype has unique properties [36].

3.4. Analysis on the BREAST Dataset. Breast cancer refers to a malignant tumor in which cancer cells have penetrated the basement membrane of breast ducts or lobular alveoli and invaded the interstitium. Many scholars have carried out a series of studies and analyses on the gene level and have given specific subtypes and treatment programs. Based on the microarray predictive analysis model, Parker et al. proposed a 50-gene classifier (known as PAM50) to classify BIC into five subtypes: basal-like, luminal A, luminal B, HER2-enriched, and normal-like [37]. On BREAST data, we counted the distribution of clustering results obtained by R^{2}SNF on the cancer subtypes basal-like, luminal A, luminal B, and HER2-enriched in Table 4. It can be seen from Table 4 that subtype 1 is mainly distributed in luminal A and luminal B (accounted for 80.6%), subtype 2 is mainly distributed in basal-like (accounted for 74.6%), and subtype 3 is mainly distributed in luminal A and luminal B (accounted for

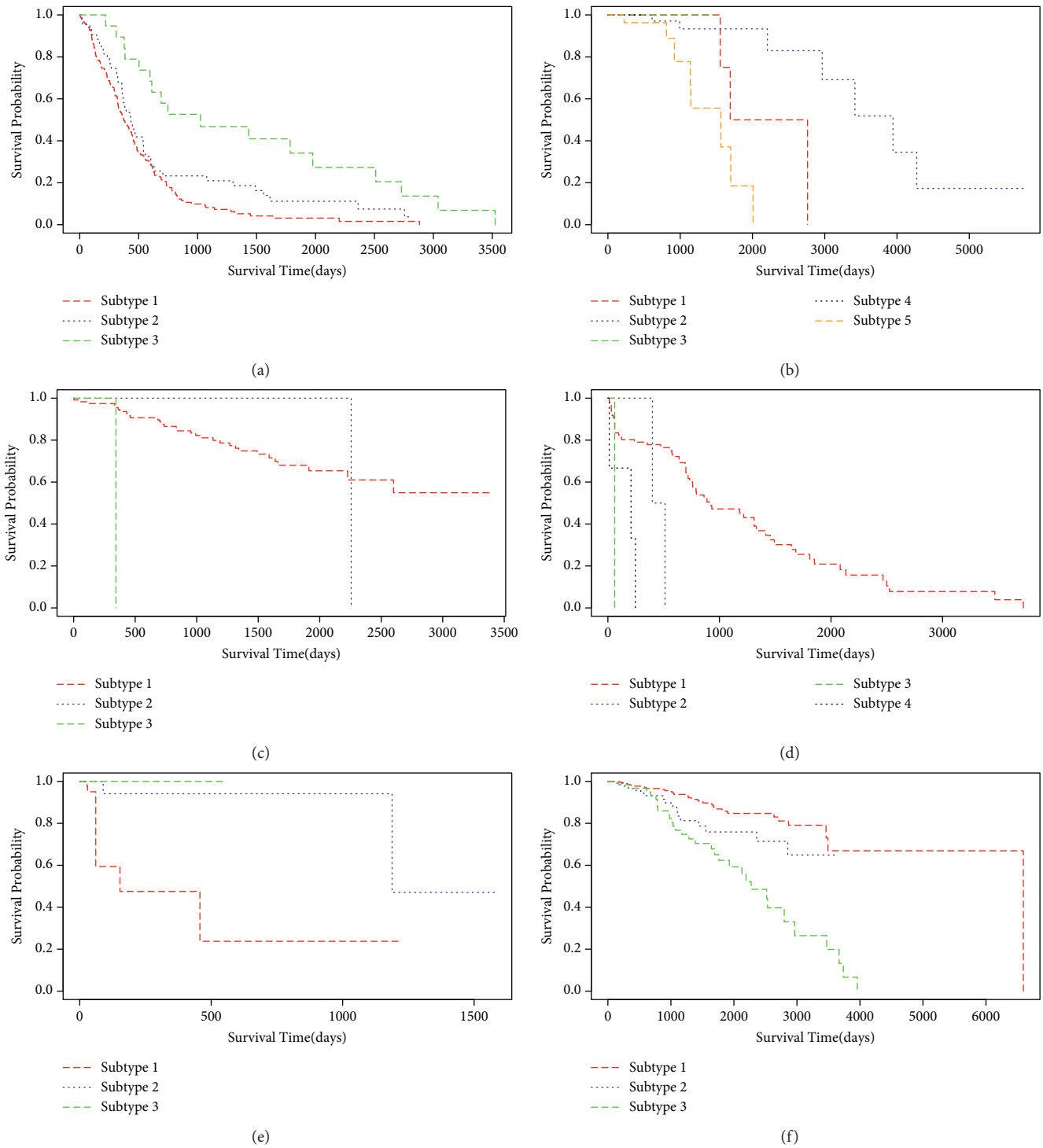
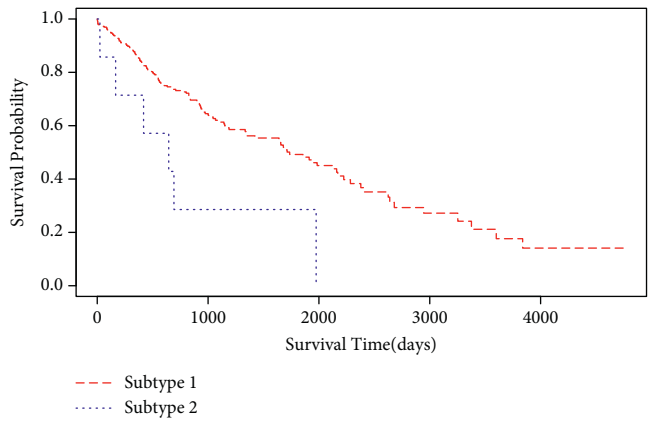


Figure 1: Continued.



(g)

Figure 1: Kaplan–Meier survival curves of different subtypes on cancer multigenomic datasets: (a) GBM, (b) BIC, (c) KRCCC, (d) LSCC, (e) COAD, (f) BREAST, and (g) LUNG.

Table 3: The distribution of subtypes obtained by R^2SNF on the subtypes determined in [34, 35].

R^2SNF subtypes	Subtypes in [34]				Subtypes in [35]	
	Mesenchymal	Classical	Neural	Proneural	G-CLMP	Non-G-CLMP
Subtype 1	46	51	26	30	1	152
Subtype 2	20	6	8	9	4	39
Subtype 3	0	0	0	19	14	5

The values in this table represent the number of patients counted.

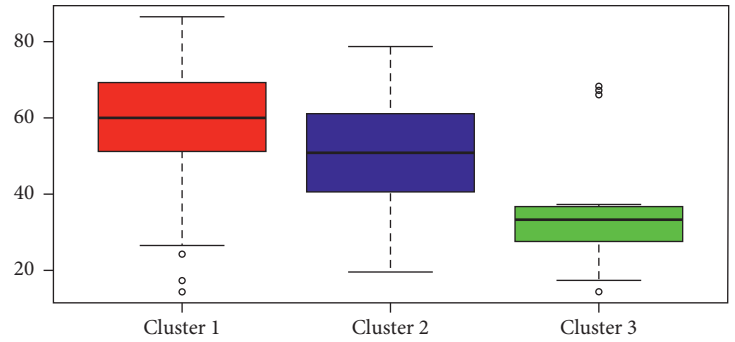


Figure 2: Boxplot of the age distribution of patients in the three cancer subtypes. The black bar represents the median of each subtype.

70.8%). In addition, we can also find that HER2-enriched is mainly distributed in subtypes 1 and 2 (accounted for 89.1%), and normal-like is mainly distributed in subtype 1 (accounted for 78.3%).

We also chose two clinical labels for which we tested enrichment: Pathologic M and Pathologic N. Pathologic M and Pathologic N are regional lymph nodes' distant metastasis stage (M) and clinical stage (N) of breast cancer, respectively. Pathologic M includes three stages: M0, M1, and MX. Pathologic N roughly includes five stages: N0, N1, N2, N3, and NX. Generally, the numbers or letters after N and M provide more details about these factors, and the higher the number, the more severe the cancer.

We used the chi-square test to verify whether there was a significant difference in our analysis among these clinical labels. The p values on Pathologic M and Pathologic N are 6×10^{-3} and 9×10^{-3}, respectively. The detailed distributions of subtypes obtained by R^2SNF on Pathologic M and Pathologic N are shown in Tables 5 and 6, respectively. In Table 5, subtype 1, subtype 2, and subtype 3 have the similar distribution: mainly distributed in M0. We calculated the proportion of samples belonging to the M0 stage in the three subtypes as 74.9%. In Table 6, subtype 1, subtype 2, and subtype 3 have the similar distribution: mainly distributed in N0 and N1. The proportion of samples belonging to the N0 stage and N1 stage in the three subtypes is 46.3% and 33.8%, respectively.

From the above analysis, we can draw the following conclusion. First, subtypes 1 and 3 are mainly distributed in luminal A and luminal B, which are the breast cancer subtypes with the best prognosis. Second, subtype 2 is mainly distributed in basal-like, in which clinical prognosis

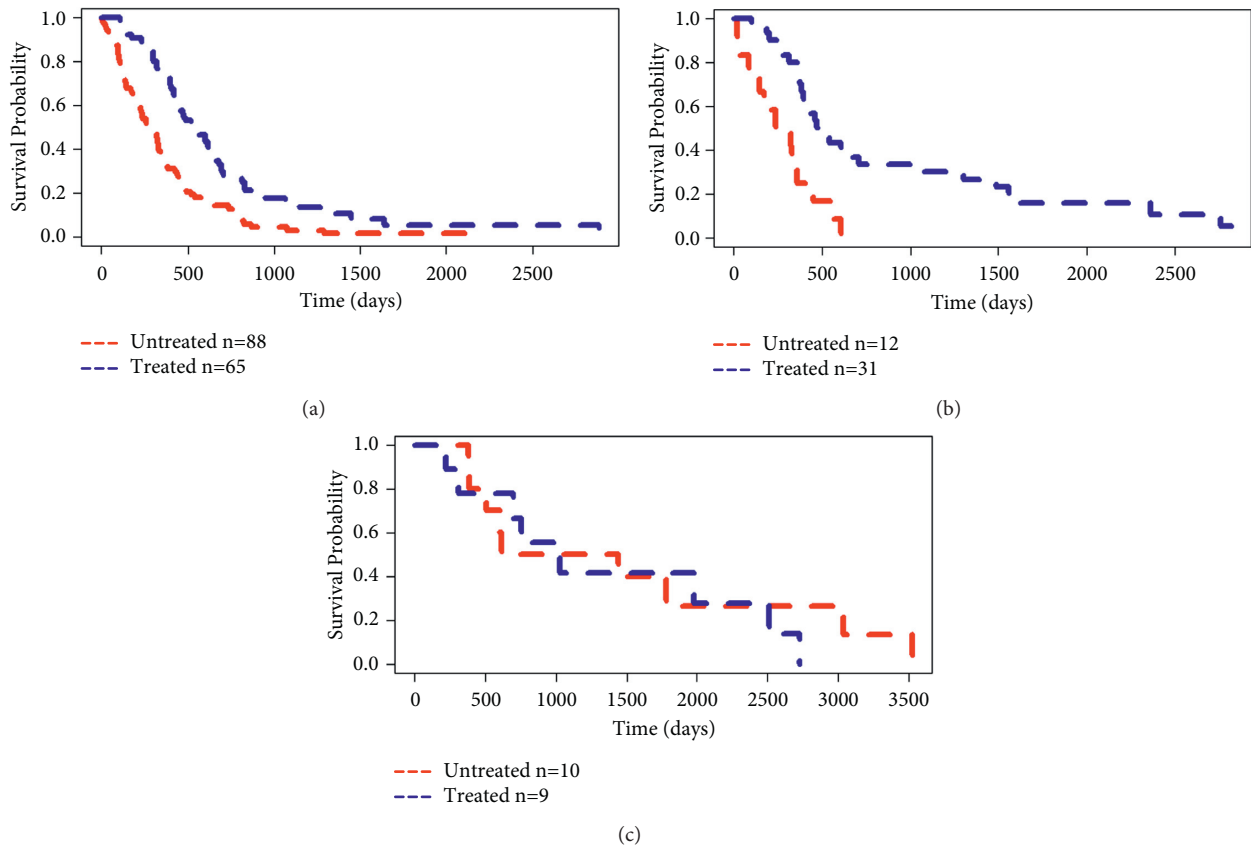


Figure 3: The Kaplan–Meier survival curves of the identified cancer subtypes by R^{2}SNF: (a) subtype 1, (b) subtype 2, and (c) subtype 3 of TMZ response. "Untreated" represents the patients who did not receive TMZ treatment, and "Treated" represents the patients who received TMZ treatment.

Table 4: The distribution of subtypes obtained by R^{2}SNF on the subtypes determined by PAM50.

R^{2}SNF subtypes	Subtypes determined by PAM50				
	Basal-like	HER2-enriched	Luminal A	Luminal B	Normal-like
Subtype 1	8	28	242	102	47
Subtype 2	97	21	2	4	6
Subtype 3	6	6	27	19	7

Table 5: The distribution of subtypes obtained by R^{2}SNF on Pathologic M.

R^{2}SNF subtypes	Pathologic M		
	M0	M1	MX
Subtype 1	305	4	118
Subtype 2	101	2	27
Subtype 3	60	1	4

Table 6: The distribution of subtypes obtained by R^{2}SNF on Pathologic N.

R^{2}SNF subtypes	Pathologic N				
	N0	N1	N2	N3	NX
Subtype 1	185	149	45	42	6
Subtype 2	79	34	13	4	0
Subtype 3	24	27	9	2	3

is poor. Third, the patients in BREAST data are mainly in the early stages of breast cancer and have high survival rate. All these conclusions can also be verified in Figure 1(f).

4. Conclusions

How to construct a robust dense similarity matrix is a key issue in SNF. In this paper, we analyzed the problems existing in the construction of the dense similarity matrix in SNF and proposed the similarity network fusion based on random walk and relative entropy (R^2SNF) method for cancer subtypes' prediction. We proposed to use the random walk with restart algorithm to characterize the complex relationship between genomic data samples and obtained the stable state transition probability distribution of each sample. We further used relative entropy to calculate the difference in the transition probability distribution between samples to construct a better dense similarity matrix which contains structural similarity information between samples. Then, the constructed dense similarity matrix and the KNN similarity matrix were nonlinearly iteratively fused. Finally, spectral clustering was used to cluster the fused similarity matrix. On seven cancer genomic datasets (GBM, BIC, KRCCC, LSCC, COAD, BREAST, and LUNG) containing three data types (mRNA expression data, miRNA expression data, and DNA methylation data), R^2SNF was compared with a variety of classical cancer subtype prediction algorithms. Experimental results show that R^2SNF has better performance in identifying cancer subtypes than the comparison algorithms. And through the analysis of the results of GBM and BREAST experiments, it can be proved that R^2SNF can discover cancer subtypes with biological significance. In addition to relative entropy, there are other methods to measure the difference between two probability distributions, such as Jensen–Shannon divergence, Wasserstein distance, and cross-entropy. In future work, we will devote ourselves to finding a more suitable method to calculate the difference between probability distributions and then to obtain a similarity matrix that is conducive to cancer subtype prediction.

Data Availability

The data used to support the findings of this study are available from the first author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 61906198, 61976215, and 61772532) and the Natural Science Foundation of Jiangsu Province (Grant no. BK20190622).

References

[1] G. Getz, S. Gabriel, K. Cibulskis et al., "Integrated genomic characterization of endometrial carcinoma," *Nature*, vol. 116, no. 7447, pp. 67–73, 2013.

[2] J. Xi, X. Yuan, M. Wang et al., "Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication," *Bioinformatics*, vol. 36, no. 6, pp. 1855–1863, 2019.

[3] C. M. Croce, "Oncogenes and cancer," *New England Journal of Medicine*, vol. 358, no. 5, pp. 502–511, 2008.

[4] A. Chen, G. Fu, Z. Xu et al., "Detection of bladder cancer via microfluidic immunoassay and single-cell DNA copy number alteration analysis of captured urinary exfoliated tumor cells," *Cancer Research*, vol. 78, no. 14, pp. 4073–4085, 2017.

[5] D. Capper, D. T. W. Jones, M. Sill et al., "DNA methylation-based classification of central nervous system tumours," *Nature*, vol. 555, no. 7697, pp. 469–474, 2018.

[6] B. Wang, A. M. Mezlini, F. Demir et al., "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[7] S. Hanash, "Integrated global profiling of cancer," *Nature Reviews Cancer*, vol. 4, no. 8, pp. 638–644, 2004.

[8] R. Shen, A. B. Olshen, M. Ladanyi et al., "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 26, no. 2, pp. 292-293, 2010.

[9] U. D. Akavia, O. Litvin, J. Kim et al., "An integrated approach to uncover drivers of cancer," *Cell*, vol. 143, no. 6, pp. 1005–1017, 2010.

[10] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 928–937, 2015.

[11] N. K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, vol. 31, no. 12, pp. i268–75, 2015.

[12] H. Wang, H. Zheng, J. Wang, C. Wang, and F.-X. Wu, "Integrating omics data with a multiplex network-based approach for the identification of cancer subtypes," *IEEE Transactions on NanoBioscience*, vol. 15, no. 4, pp. 335–342, 2016.

[13] T. L. Van, L. M. Van, and A. C. Fierro, "Simultaneous discovery of cancer subtypes and subtype features by molecular data integration," *Bioinformatics*, vol. 32, no. 17, pp. 445–454, 2016.

[14] E. Zhang and X. Ma, "Regularized multi-view subspace clustering for common modules across cancer stages," *Molecules*, vol. 23, no. 5, p. 1016, 2018.

[15] M. Hofree, J. P. Shen, H. Carter et al., "Network-based stratification of tumor mutations," *Nature Methods*, vol. 10, no. 11, pp. 1108–1115, 2014.

[16] J. K. Huang, T. Jia, D. E. Carlin, and T Ideker, "pyNBS: a Python implementation for network-based stratification of tumor mutations," *Bioinformatics*, vol. 34, no. 16, pp. 2859–2861, 2018.

[17] T. Xu, L. T. Duy, L. Lin, R. Wang, B. Sun, and J. Li, "Identifying cancer subtypes from miRNA-TF-mRNA regulatory networks and expression data," *PloS One*, vol. 11, no. 4, Article ID e0152792, 2016.

[18] C. Yang, S.-G. Ge, and C.-H. Zheng, "ndmaSNF: cancer subtype discovery based on integrative framework assisted by network diffusion model," *Oncotarget*, vol. 8, no. 51, pp. 89021–89032, 2017.

[19] B. Yang, Y. Zhang, S. Pang, X. Shang, X. Zhao, and M. Han, "Integrating multi-omic data with deep subspace fusion clustering for cancer subtype prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 216–226, 2019.
[20] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
[21] K. Pearson, "The problem of the random walk," *Nature*, vol. 72, no. 1865, p. 294, 1905.
[22] H. Tong, C. Faloutsos, J. Pan et al., "Fast random walk with restart and its applications," in *Proceedings of the International Conference on Data Mining*, pp. 613–622, Washington, DC, USA, December 2006.
[23] V. Martinez, F. Berzal, and J. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, vol. 49, no. 4, p. 69, 2017.
[24] W. Zheng, S. Liu, and J. Mu, "A random walk similarity measure model based on relative entropy," *Journal of Nanjing University (Natural Science)*, vol. 55, no. 6, pp. 984–999, 2019.
[25] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
[26] N. Rappoport and R. Shamir, "Multi-omic and multi-view clustering algorithms: review and cancer benchmark," *Nucleic Acids Research*, vol. 46, no. 20, pp. 10546–10562, 2018.
[27] K. Akazawa, T. Nakamura, and Y. Palesch, "Power of logrank test and cox regression model in clinical trials with heterogeneous samples," *Statistics in Medicine*, vol. 16, no. 5, pp. 583–597, 1997.
[28] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.
[29] D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification," *BMC Genomics*, vol. 16, no. 1, p. 1022, 2015.
[30] Q. Mo, S. Wang, V. E. Seshan et al., "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.
[31] Q. Shi, C. Zhang, M. Peng et al., "Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data," *Bioinformatics*, vol. 33, no. 17, pp. 2706–2714, 2017.
[32] T. Ma and A. Zhang, "Affinity network fusion and semi-supervised learning for cancer patient clustering," *Methods*, vol. 145, pp. 16–24, 2018.
[33] J. Tian, J. Zhao, and C. Zheng, "Clustering of cancer data based on Stiefel manifold for multiple views," *BMC Bioinformatics*, vol. 22, no. 1, p. 268, 2021.
[34] R. G. W. Verhaak, K. A. Hoadley, E. Purdom et al., "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.
[35] H. Noushmehr, D. J. Weisenberger, K. Diefes et al., "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, vol. 17, no. 5, pp. 510–522, 2010.
[36] C. W. Brennan, R. G. Verhaak, A. McKenna et al., "The somatic genomic landscape of glioblastoma," *Cell*, vol. 155, no. 2, pp. 462–477, 2013.
[37] J. S. Parker, M. Mullins, M. C. U. Cheang et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, 2009.