A Clustering Algorithm for Tumor Gene Data Based on Improved DPC Algorithm

Wei Wang, Bo Gao*

Computer Department
Qinhuangdao Radio and Television University
Qinhuangdao 066000, China
E-mails: qhdwang2020@163.com, gabosky@163.com

**Corresponding author*

Received: September 10, 2021 **Accepted: March 17, 2022**

Published: June 30, 2022

Abstract: *Cluster analysis is a principal approach to discover unknown tumor subtypes. Innovative and effective cluster analysis methods are of great significance for tumor diagnosis and malignant tumor treatment. Existing studies on the cluster analysis of tumor gene data generally have defects in aspects such as unsatisfactory performance in clustering high-dimensional and high-noise data, and insufficient accuracy in selecting cluster centers. To overcome these defects, this paper performed cluster analysis on tumor gene data based on an improved Density peaks clustering (DPC) algorithm. At first, this paper elaborated on the composition and storage format of tumor tissue samples used in the experiment, gave the tumor gene expression profile data in the matrix format, and introduced the preprocessing process of gene expression profile data. Then, this paper carried out feature selection of tumor gene expression profile data. At last, this paper innovatively divided the target gene density into two parts of K-nearest neighbor local density and neighborhood density, thereby completing the improvement of conventional DPC algorithm and expanding its application scenarios. Combining with experiment, the clustering results of the algorithm before and after introducing the idea of Approximate Nearest Neighbor (ANN) were given, which had verified the effectiveness of the algorithm proposed in this paper.*

Keywords: *Cluster analysis, Tumor genes, DPC algorithm, Feature selection, Gene density.*

Introduction

When the chromosomal DNAs inside human body cells are damaged under the action of various carcinogenic factors and undergo gene mutations, local tissue cells would proliferate abnormally and form tumors [1, 10, 13, 15, 25]. Tumors have many subtypes; as a heterogeneous disease, there are also many expression patterns of tumor genes [7, 11, 12, 26]. Therefore, in terms of tumor gene data, the traditional research methods based on single gene usually have great limitations [3, 5, 9, 16, 23, 27, 31]. Gene chip technology can simultaneously process large-scale biological information, making it possible to compare and analyze gene expression data under both normal and diseased states [2, 4, 14, 17-21, 28, 30]. At current stage, cluster analysis is a principal approach to discover unknown tumor subtypes. Innovative and effective cluster analysis methods are of great significance for tumor diagnosis and malignant tumor treatment.

As early as in the 1980s, the subtypes of tumors have been validated theoretically and practically, but the classification is relatively simple. Qaddoum [22] introduced an advanced method for classifying tumor types through microarray gene selection records, using shuffling-based gene selection and optimized data clustering. Qaddoum developed a new hybrid algorithm that combines the artificial bee colony algorithm with genetic algorithm, and took it as the clustering tool of key gene selection. Bladder cancer is a common urinary system tumor with a higher incidence in men between 60 and 70 years old. Sarafidis et al. [24] employed bioinformatics analysis and regression method to find common gene expression profiles related to tumor subtypes and differentiation, their method was proved to be helpful for determining noel gene targets, which can be used as targets for prognosis, diagnosis, and treatment. High-dimensional data such as the data of gene expression profiles generally have high homogeneity and high noise, and there's a large amount of redundant information in the data in the same database. Dai et al. [6] proposed a new non-negative matrix factorization algorithm called the sparse orthogonal non-negative matrix factorization; they applied it to identify differentially expressed genes and cluster tumor samples, added L1 norm regularization and orthogonal constraints to the traditional Non-Negative Matrix Factorization (NMF) model to obtain a more powerful data analysis tool. Existing studies have proposed different algorithms for tumor clustering, but few of them made use of the knowledge of experts to improve the performance of tumor discovery. Yu et al. [29] took expert knowledge as a constraint in the clustering process, and proposed a semi-supervised clustering ensemble framework based on feature selection, and applied it to tumor clustering of biomolecular data. In order to evaluate the robustness of using sparse manifold clustering and embedding to classify gene expression profiles, García-Gómez et al. [8] adopted Sparse Manifold Clustering and Embedding (SMCE) algorithm to reduce the dimensionality of preprocessing dataset, and used both supervised and unsupervised methods to obtain the classification model: the former method was based on linear discrimination analysis, and the latter performed clustering based on SMCE embedded data.

After reviewing and summarizing above references, it is found that world field scholars have achieved certain research results in terms of the cluster analysis of tumor gene data, but there are still shortcomings. Some algorithms have poor performance in clustering high-dimensional and high-noise data, the correction rate of cluster center selection is unsatisfactory, which has resulted in insufficient accuracy of clustering results, and even failed to produce biologically meaningful explanations. In view of these defects, this study performed cluster analysis on tumor gene data based on an improved Density peaks clustering (DPC) algorithm, and the main content of this paper contains these aspects: 1) introduce in detail the composition and storage format of tumor tissue samples used in the research; 2) give the tumor gene expression profile data in the matrix form, and explained the preprocessing process of gene expression profile data; 3) complete the feature selection of tumor gene expression profile data; 4) divide the target gene density into two parts: the *K*-nearest neighbor local density and neighborhood density, and complete the improvement of the traditional DPC algorithm, give a detailed description of the execution steps of the proposed improved DPC algorithm, and use experimental results to verify the effectiveness of the proposed algorithm.

The research conducted in this paper designed an effective cluster algorithm for gene expression profile data, in the hopes of providing valuable reference opinions for molecular therapy, targeted drug recommendation, and cancer prediction.

Dataset structure and preprocessing method

Tumor is a multi-gene, multi-factor disease; the tumor gene expression profile data has the characteristics of containing lots of noise data and redundant data, and the data are often of high dimensions. This paper selected two classic tumor gene expression profile datasets for research: the acute leukemia dataset, and the digestive system tumor dataset.

Tissue samples studied in this paper included: 48 acute lymphocytic leukemia *ALL*, 22 acute myeloid leukemia *AML*, 18 chronic lymphocytic leukemia *CLL*, and 26 chronic myelogenous leukemia *CML*. A gene expression matrix was built based on 1827 gene expression data. Table 1 gives the storage format of some gene expression data.

Table 1. Storage format of acute leukemia gene expression data

Probe number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Description	**Control sequence**								
*ALL*1	-163.7	35.6	-33.2	176.2	117.7	15	18.5	-52.7	23.5
*ALL*2	-236	-105	-32	62	158	-34	165	-7	-35
*AML*1	-287	93	-32	145	-2	22	286	-115	93
*AML*2	-21	116	-12	27	182	31	304	-72	-28
*CLL*1	-265	129	82	338	145	248	207	-313	0
*CML*1	-374	11	23	156	85	118	339	-7	75

According to the difference of epigenetic genes, gastric cancer is divided into proliferative type *PRO*, metabolic type *MET*, and interstitial type *INT*.

The dataset used in this research came from the gastric cancer tumor expression profile dataset of the GEO database of the National Center for Biotechnology Information of the United States, it contains expression profiles of 21 samples, and expression data of 475 genes. The storage format of some gene expression data is given in Table 2.

Table 2. Storage format of gastric cancer gene expression data

Probe number	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Gene name	*SEMA*3*C*	*GML*	*MKNK*1	*OGG*1	*FAM*193*A*	*SH*3*BP*2	*C*4*orf*10
*GSM*51763	49.2	3	75.6	11.5	205.3	31.6	86.4
*GSM*51764	203.4	31.2	33.6	137.4	215.7	23.5.	64.7
*GSM*51765	82.6	1.8	157.4	9.6	286.4	13.7	72.5
*GSM*51766	68.3	1.4	164.8	8.5	218.3	11.4	73.5
*GSM*51767	39.7	12.3	101.9	14.5	241.2	7.8	79
*GSM*51768	44.6	4.5	236.5	12.4	354.7	28.9	136.8
Probe number	**8**	**9**	**10**	**11**	**12**	**13**	
ID EN TIFIER	*GABRA*3	*OMD*	*IFI*44*L*	*VRK*1	*VRK*2	*C*4*orf*10	
*GSM*51763	23.8	7.8	37.6	19.6	31.7	10	
*GSM*51764	44.8	9.5	73.6	32.4	67.9	8.6	
*GSM*51765	15.8	3.6	202.2	22.7	42.5	12.2	
*GSM*51766	27.8	16.4	31.7	25.9	19.3	11.8	
*GSM*51767	45.3	73	35.5	26.6	34.2	7.6	
*GSM*51768	87	43.5	138.5	54.3	105.9	17.9	

The preprocessing of gene expression profile data is the key to subsequent cluster analysis of tumor gene data, and it can be performed in three steps: data cleaning, data filling, and data normalization.

Fig. 1 gives the matrix form of gene expression profile data. Data cleaning is to delete meaningless data elements in the matrix. In a complete gene sample, if the total number of missing features i is lower than threshold ψ, then the feature will be deleted before subsequent cluster analysis, usually the value of ψ takes 3.

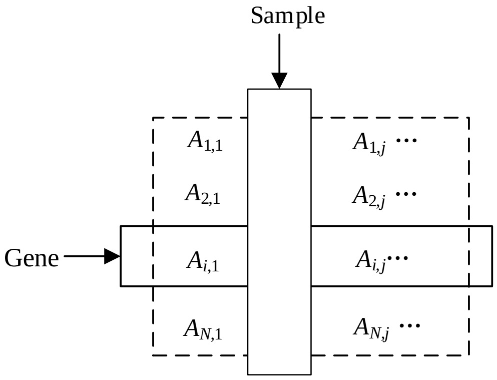


Fig. 1 The matrix form of gene expression profile data

Directly deleting the missing items in the matrix may result in loss of valuable features. As the primary preprocessing method of tumor gene expression data, data filling needs to adopt a certain filling strategy to give more complete gene expression profile datasets. This paper chose to use the *K*-nearest neighbor filling method to fill in the missing values, that is, select *K* neighbor genes that are closest to the missing item in the gene expression profile data, calculate the corresponding *K* distances, determine the weight values of the *K* neighbor genes based on the distance, and fill in the missing items based on the weighting method. Suppose: δ_i represents the Euclidean distance between the missing items in the gene expression profile data and the *i*-th neighbor gene; ω_i represents the weight of the *i*-th neighbor gene, then Eq. (1) gives the expression of weight value:

$$\omega_i = \frac{\delta_i^{-1}}{\sum_{i=1}^{L} \delta_i^{-1}} \tag{1}$$

The filling values of missing items in gene expression profile data can be obtained from Eq. (2):

$$MV = \sum_{i=1}^{L} \omega_i \times H\_EX_i \tag{2}$$

By normalizing the gene expression profile dataset, the value range of all data in the dataset could be adjusted to the [0, 1] interval, in this way, the gene expression profile data of different dimensions become comparable. Suppose: a_{ij} represents the original gene expression value; *A* represents the gene expression value after normalization; $a_{\min}$ and $a_{\max}$ respectively represent the minimum and maximum values of gene expression in the sample, then Eq. (3) gives the calculation formula:

$$A = \frac{a_{ij} - a_{\min}}{a_{\max} - a_{\min}} \tag{3}$$

In order to further eliminate the influence of different data caused by mutations, this paper normalized the standard deviation of the gene expression profile dataset. The normalization is based on the original gene expression value *A*, the mean value A' and standard deviation ε were normalized, and Eq. (4) gives the calculation formula:

$$C - S = \frac{A - A'}{\varepsilon} \tag{4}$$

After subjected to the standard deviation normalization, the gene expression profile dataset obeys the standard normal distribution, namely the normal distribution with 0 as the mean and 1 as the standard deviation.

Feature selection of tumor gene expression profile data

To reduce the deviation of results brought by the cluster analysis of tumor gene data, it is necessary to solve the "curse of dimensionality" caused by the characteristics of tumor gene expression profile data. The number of decisive genes with significant tumor-causing characteristics in the sample genes is less than the feature dimension of the sample gene expression profile dataset. In order to effectively improve the accuracy of tumor gene data clustering, it needs to screen out a key feature subset from the sample gene expression profile dataset for modeling and completing data feature selection.

Suppose: A_i represents the quantitative value of the expression amount of gene i in a normal sample after subjected to normal normalization transformation, it obeys the normal distribution with a mean of n_i and a variance of m_i^2, and $n_i = 0$, $m_i^2 = 1$, namely $A_i \sim M(0, 1)$. Suppose: B_i represents the value of pure tumor gene i after subjected to the same normal normalization transformation, then $B_i = A_i + \xi_i$, wherein ξ_i represents the difference in the quantitative values of the expression amount of gene i in pure tumor samples and in normal non-diseased samples. If ξ_i also obeys normal distribution and satisfies $\xi_i \sim M(o_i, w_i^2)$, then B_i also obeys normal distribution and satisfies $B_i \sim M(o_i + 0, w_i^2 + 1)$. The difference in the quantitative values of gene expression amount can be represented by $A_i - B_i$, and the gene feature selection problem can be transformed into a hypothesis test problem of F_0: $v_i = 0$.

In fact, in the cluster analysis of tumor gene data, the adopted gene expression profile data are usually the combinations of tumor samples and normal samples. Suppose: B_i^* represents the quantitative value of gene expression amount, γ_{AC} represents the tumor purity of tumor sample AY, then the quantitative value of gene expression amount of real samples could be determined based on the idea of linearity hypothesis:

$$\begin{aligned} B_i^* &= (1-\gamma_{AC})A_i + \gamma_{AC}B_i \\ &= (1-\gamma_{AC})A_i + \gamma_{AC}(A_i + \xi_i) = A_i + \gamma_{AC}\xi_i \end{aligned} \tag{5}$$

According to the Eq. (5), B_i^* also obeys normal distribution and satisfies

$B_i^* \sim M(\gamma_{AC}o_i, 1 + \gamma^2{}_{AC}w^2{}_i)$

γ_{AC} can be taken as a covariate and added into *DESeq*2, the statistical tool for differential gene screening, to carry out test and analysis on the differentially expressed genes.

Compared with other gene difference analysis methods such as the method of variance analysis, the fold change analysis method is simpler, which only needs to compare the gene data analysis results of experimental group and control group based on fluorescence intensity. The gene expression profile data of tumor gene data cluster analysis is composed of two parts: the amount of tumor gene expression, and the amount of normal gene expression. Suppose: a_{CG} represents the amount of tumor gene expression, a_{NG} represents the amount of normal gene expression, the mean(·) function is a function used to calculate the average value, η_D represents the ratio of the mean of a_{CG} to the mean of a_{NG}, then, the calculation formula of fold change analysis is given by Eq. (6):

$$\eta\_D = \frac{mean(a_{CG})}{mean(a_{NG})} \tag{6}$$

The value of η_D usually satisfies $|\log_2 \eta\_D| > 1$. Although this method is simple and intuitive, it does not fully consider the statistical significance of gene difference quantification. Furthermore, this paper adopted T-test to measure the differential expression of genes from a statistical point of view. Suppose: $VA^2{}_{CG}$ and $VA^2{}_{NG}$ respectively represent the variance of tumor gene samples and the variance of normal gene samples, M_1 and M_2 respectively represent the number of tumor gene samples and the number of normal gene samples, and they obey the t distribution with a degree of freedom of $M_1 + M_2 - 2$, then the calculation formula of T statistics could be expressed as Eq. (7):

$$\tau_{gen\_data} = \frac{mean(a_{ill}) - mean(a_{regular})}{\sqrt{\frac{VA_{ill}^2}{M_1} + \frac{VA_{regular}^2}{M_2}}} \tag{7}$$

The improved DPC clustering algorithm for tumor gene expression profile data

The tumor gene expression profile data including tumor gene expression and normal gene expression has the characteristics of uneven density distribution and intertwined, the traditional DPC density peak clustering algorithm is not applicable for these characteristics, so the clustering performance is poor. For this reason, this paper aimed to improve the traditional DPC algorithm, it divided the target gene density into two parts: K-nearest neighbor local density, and neighborhood density, and the neighborhood density was obtained by calculating the ratio of the local density of gene expression profile data to its K-nearest neighbor local density and distance.

For a given gene expression profile dataset $A = \{a_1, a_2, ..., a_m\}$, suppose a_j represents the K-nearest neighbors of the target gene a_i, δ_{ij} represents the Euclidean distance between target genes a_i and a_j, then, the local density σ_{i1} of a_i can be calculated by Eq. (8):

$$\sigma_{i1} = \sum_{j \in LMM_i} r^{-\delta_{ij}} \tag{8}$$

Suppose: σ_{i2} represents the neighborhood density ratio of each target gene a_i, then its value can be calculated by Eq. (9):

$$\sigma_{i2} = \sum_{j \in LMM_i} \frac{\sigma_{i1}}{\sigma_{j1} \times (\delta_{ij} + 1)} \tag{9}$$

Eq. (10) gives the calculation formula of the real density of target gene a_i:

$$\sigma_i = \sigma_{i1} + \sigma_{i2} \tag{10}$$

By superimposing the *K*-nearest neighbor local density and neighborhood density, and calculating the ratio of the local density of target gene to its neighborhood density, it can effectively avoid the situation that the sum of *K*-nearest neighbors in the tumor gene expression profile data might is the same, thereby truthfully reflecting the density of each target gene in the tumor gene expression profile dataset.

For a given gene expression profile dataset $A = \{a_1, a_2, ..., a_m\}$, suppose IN_i represents the inverse nearest neighbor set of target gene a_i, KN_i represents the set of *K*-nearest neighbors, then the inverse nearest neighbor of a_i can be defined by Eq. (11):

$$IN_i = \{ j \in A, i \in KN_j \} \tag{11}$$

According to above formula, the influence on target genes in the tumor gene expression profile dataset can be described by the inverse nearest neighbor set of each original target gene. When a target gene is located in a high-density region, it is usually surrounded by many other genes. When a target gene is located in a low-density region, it means that the target gene has few neighbor genes.

The influence space $YK(i)$ of target gene a_i can be defined by Eq. (12):

$$YK(i) = KN_i \cap IN_i \tag{12}$$

The influence space *YK* can describe the two-way neighborhood relationship between target genes. The tightness between two target genes described by the one-way *K*-nearest neighbor is less than the tightness between two target genes within the influence space, that is, the accuracy of similarity between the two target genes in the influence space is higher.

The number of shared *K*-nearest neighbors of any two target genes a_i and a_j can be calculated by Eq. (13):

$$SKN_{ij} = KN_i \cap KN_j \tag{13}$$

For the similarity between target genes a_i and a_j in a given gene expression profile dataset $A = \{a_1, a_2, ..., a_m\}$, and whether there is a shared *K*-nearest neighbor between the two and its influence space, this paper classified target genes a_i and a_j and gave the definition of their similarity in different situations.

1) In case that target genes a_i and a_j have influence space and shared *K*-nearest neighbors at the same time, their similarity can be calculated by Eq. (14):

$$TGS_{ij} = 2 \times r^{-\delta_{ij}} + r^{-\frac{\delta_{ij}}{EMM_{ij}}} + \frac{min(\sigma_i, \sigma_j)}{max(\sigma_i, \sigma_j)} \tag{14}$$

2) In case that target genes a_i and a_j do not have shared *K*-nearest neighbors but have influence space, their similarity can be calculated by Eq. (15):

$$TGS_{ij} = 2 \times r^{-\delta_{ij}} + \frac{min(\sigma_i, \sigma_j)}{max(\sigma_i, \sigma_j)} \tag{15}$$

3) In case that target genes a_i and a_j have shared K-nearest neighbors but do not have influence space, and a_j is the K-nearest neighbor of a_i, their similarity can be calculated by Eq. (16):

$$TGS_{ij} = r^{-\delta_{ij}} + r^{-\frac{\delta_{ij}}{EMM_{ij}}} + \frac{min(\sigma_i,\sigma_j)}{max(\sigma_i,\sigma_j)} \tag{16}$$

4) In case that target genes a_i and a_j have shared K-nearest neighbors but do not have influence space, and a_j is not the K-nearest neighbor of a_i, their similarity can be calculated by Eq. (17):

$$TGS_{ij} = r^{-\frac{\delta_{ij}}{EMM_{ij}}} + \frac{min(\sigma_i,\sigma_j)}{max(\sigma_i,\sigma_j)} \tag{17}$$

5) In case that target genes a_i and a_j do not have influence space or shared K-nearest neighbors, but a_j is the K-nearest neighbor of a_i, their similarity can be calculated by Eq. (18):

$$TGS_{ij} = r^{-\delta_{ij}} + \frac{min(\sigma_i,\sigma_j)}{max(\sigma_i,\sigma_j)} \tag{18}$$

6) In case that target genes a_i and a_j do not have influence space or shared K-nearest neighbors, and a_j is not the K-nearest neighbor of a_i, their similarity can be calculated by Eq. (19):

$$TGS_{ij} = 0 \tag{19}$$

For target gene x_ia_i, its similarity KL-nearest neighbor set is the KL target genes with the greatest similarity with x_ia_i, at the same time, the KL-nearest neighbor set was replaced:

$$IN_i \to TGS(IN_i) \tag{20}$$

Besides making the targe gene data easy to fall in higher local density, the DPC algorithm's target gene allocation strategy also needs to be improved. Fig. 2 shows the similarity of density distribution between target genes.

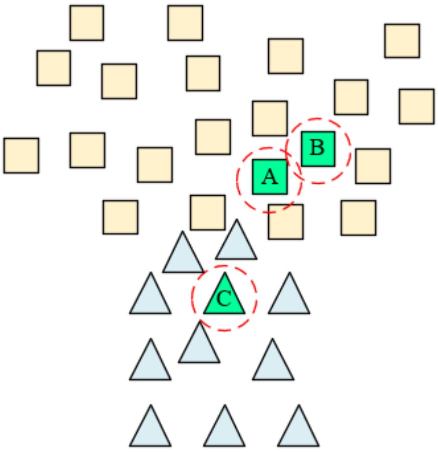


Fig. 2 A diagram of density distribution similarity between target genes

In order to effectively improve the grade of membership between samples and the accuracy of allocation results, this paper had fully considered both the distance similarity and density distribution similarity at the same time, and optimized the cluster expansion and allocation strategy of the algorithm.

For each un-allocated gene a_i, a_j is a similar neighbor of a_i, and a_j has been allocated already, then the label of the class that a_j has been allocated to can be described by $b_j = d$ ($d = 1, 2, ..., n$), and there are $B_{ij} = q_{ij}/(\Sigma_{k \in RES\ LMMi} q_{ij})$, $q_{ij} = 1/(1 + \delta_{ij})$. Suppose: Φ_{ij} represents the normalized value of similarity, $|\sigma_i - \sigma_j|$ represents the density difference, and then Eq. (21) gives the calculation formula of the grade of membership OW:

$$OW_i^d = \sum_{b_i = d}^{j \in TGS[IN(i)]} \Phi_{ij} \times q_{ij} \times \left[\frac{1}{|\sigma_i - \sigma_j| + 1} \right] \tag{21}$$

According to above formula, OW is affected by two influencing factors: the $TGS[IN(i)]$ of targe genes, and the density difference between two target genes, that is, its size is jointly determined by Φ_{ij} and density difference. Greater Φ_{ij} value means that the distance between two target genes is closer; a smaller density difference means that the distribution similarity of two target genes is higher. The former case conforms to the situation that closer clusters are of higher similarity, and the latter case conforms to the situation that clusters of similar data distribution are of higher similarity. The smaller the Φ_{ij} and density difference values, the smaller the influence of allocated target gene a_j on the grade of membership of un-allocated gene a_i, and the greater the probability of two genes belonging to a same cluster.

Through the above method, the un-allocated gene a_i can be allocated to the cluster with the largest OW, thereby realizing effective improvement of the allocation strategy of the DPC algorithm.

Based on the clustering expansion process of tumor gene data, it can be known that, two gene clusters with intersection may merge into one cluster under the influence of intersection. Therefore, before allocating the rest target genes, it is necessary to complete the labeling of low-cohesion points, suppose α represents the cohesion of target gene a_j, then Eq. (22) gives its definition:

$$\alpha_j^l = \frac{l}{\sum_{k \in LMM_i} \delta_{jl}} \tag{22}$$

The mean of the sum of cohesion values of all genes is defined as the standard value of cohesion, which is represented by γ, then there is:

$$\gamma = \frac{1}{M} \sum_{j=1}^{M} \alpha_j^l \tag{23}$$

Gene datasets below the standard value of cohesion are defined as low-cohesion sets, then there is:

$$COH = \{\alpha_j^l < \gamma\} \tag{24}$$

The specific steps of the DPC algorithm are (Fig. 3).

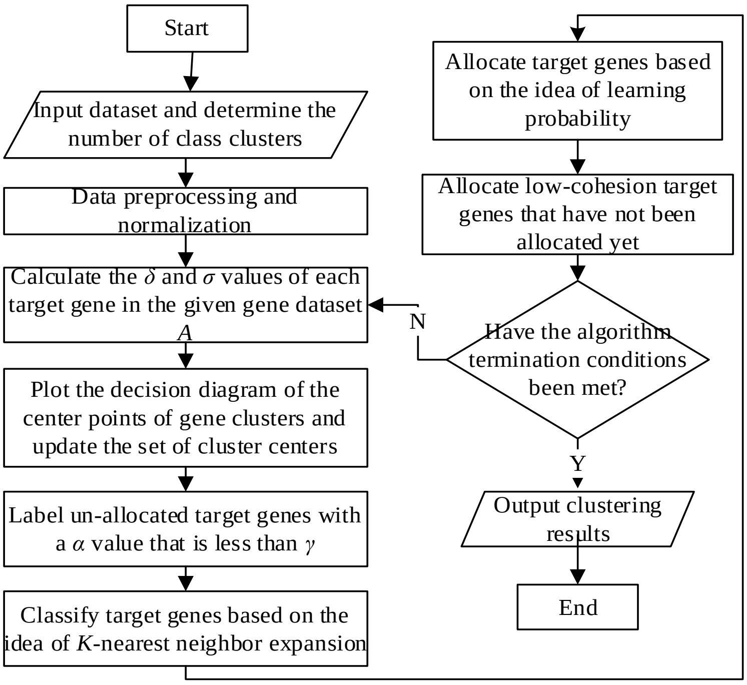

Start
Input dataset and determine the number of class clusters
Data preprocessing and normalization
Calculate the δ and σ values of each target gene in the given gene dataset A
Plot the decision diagram of the center points of gene clusters and update the set of cluster centers
Label un-allocated target genes with a α value that is less than γ
Classify target genes based on the idea of K-nearest neighbor expansion
Allocate target genes based on the idea of learning probability
Allocate low-cohesion target genes that have not been allocated yet
Have the algorithm termination conditions been met?
N
Y
Output clustering results
End

Fig. 3 Execution flow of the improved DPC algorithm

Step 1: Perform preprocessing on gene expression profile data adopted for the cluster analysis of tumor gene data, such as data cleaning, data filling, and data normalization.

Step 2: According to the definition of the real density of target gene a_i, calculate the Euclidean distance δ between each target gene in the given gene dataset A and other genes, and the local density σ.

Step 3: Plot the decision diagram of the center points of gene clusters based on the δ and σ values of each target gene, the center points of gene clusters are points with greater decision values, then, the selected cluster center points are added into the set of cluster centers.

Step 4: Calculate $TGS[IN(i)]$.

Step 5: Label each un-allocated target gene j in the updated cluster center set as "allocated", and add the $TGS[IN(i)]$ of cluster to which j is allocated into the initialized sequence DL.

Step 6: Calculate α (the cohesion value of each gene) and γ (the standard value of the cohesion of all genes), and label the un-allocated target genes with an α value smaller than γ.

Step 7: Classify target genes based on the idea of K-nearest neighbor expansion:

1) Select the first target gene i in sequence DL, allocate target gene i satisfying the similarity value conditions to the cluster it belongs, and add it to the tail of sequence DL. The similarity value conditions are: satisfies $TGS_{io} > mean(TGS_{ol})$, un-allocated, and has high cohesion $TGS[KN]$, wherein $o \in TGS[KN]_i$ and $l \in TGS[KN]_o$.

2) If sequence DL is not empty, go to **Step 7-1**.

Step 8: Allocate target genes based on the idea of learning probability:

1) Build an *m*×*n* dimensional matrix *TGM* with the *m* un-allocated target genes in **Step 7**, calculate and store the grade of membership OW_m^d of each un-allocated target gene; then screen elements in each row of matrix *TGM*, pick the gene corresponding to the maximum value of the grade of membership *max*(OW_m^d) and add it to gene list *LO*, and then add the cluster category *argmax*(OW_m^d) of the gene to list *LMO*.

2) If there are still un-allocated target genes, then search and allocate *argmax*(OW_i^d). If there's no un-allocated target gene, then go to **Step 8-3**.

3) Update matrix *TGM* and lists *LO* and *LMO*, and update the grade of membership of each target gene *w* in *TGS*[*KN*] according to the formula below:

$$OW_w^d = OW_w^d + \Phi_{iw} \times q_{iw} \times \left[\frac{1}{|\sigma_i \times \sigma_w| + 1} \right] \tag{25}$$

Update *max*(OW_m^d) and *argmax*(OW_m^d) corresponding to list *LO* and list *LMO*, and go to **Step 8-2**.

Step 9: If there are still some low-cohesion target genes that have not been allocated yet, then allocate them to the nearest cluster according to the principle of proximity.

Experimental results and analysis

This paper selected two clustering performance evaluation indexes, the Adjusted Rand Index (ARI) and the F-1 value, to analyze the clustering results of tumor gene data. The greater the ARI and F-1 value, the closer the experimental results are to the real situation. Tablec 3 and 4, respectively give the ARI and F-1 value of the clustering results of tumor gene data. For the proposed algorithm, during the clustering process of 10 listed gene samples, the values of *K*-nearest neighbor local density and neighborhood density had an impact on the grade of membership, thereby affecting the clustering performance of the algorithm.

Table 3. ARI values

Algorithm		**Traditional DPC**	**Introduce the idea of similar neighbors**	**The proposed algorithm**
Serial number of the sample	**1**	0.712	0.832	0.765
	2	0.765	0.886	0.883
	3	0.785	0.984	1.000
	4	0.831	0.735	0.846
	5	0.865	0.923	0.962
	6	0.867	0.823	0.876
	7	0.524	0.514	0.621
	8	0.368	0.495	0.532
	9	-0.023	0.386	0.758
	10	0.542	0.647	0.658

According to the above two tables, in most gene samples, the proposed algorithm had greater ARI and F-1 values, therefore, in terms of the overall performance, the proposed algorithm is better than the other two algorithms.

Table 4. F-1 values

Algorithm		**Traditional DPC**	**Introduce the idea of similar neighbors**	**The proposed algorithm**
Serial number of the sample	**1**	0.752	0.826	0.863
	2	0.831	0.958	0.968
	3	0.765	0.984	1.000
	4	0.946	0.872	0.954
	5	0.971	0.945	0.972
	6	0.993	0.954	0.996
	7	0.845	0.821	0.962
	8	0.516	0.838	0.839
	9	0.457	0.685	0.876
	10	0.924	0.910	0.945

Fig. 4 compares the similarity of different algorithms. In terms of the proposed algorithm, according to different situations of whether there're shared *K*-nearest neighbors and influence space or not, different similarity values could be obtained. Through the algorithm comparison experiment of three cases with the greatest possibility, it can be seen that, for different experiment samples, their optimal similarity values were different, the greater the similarity, the greater the grade of membership of the gene, and the greater the influence on the cluster centers that distinguish a same class and other genes. Therefore, the model proposed in this paper has certain reference value for the cluster analysis of real tumor gene data.

For the proposed algorithm, a decision diagram of the center points of gene clusters was plotted based on Euclidean distance δ and local density σ of target genes and other genes, target genes with greater δ and σ values were selected as the cluster centers. Fig. 5 shows the distribution of target genes in a two-dimensional space arranged in the order of decreasing local density.

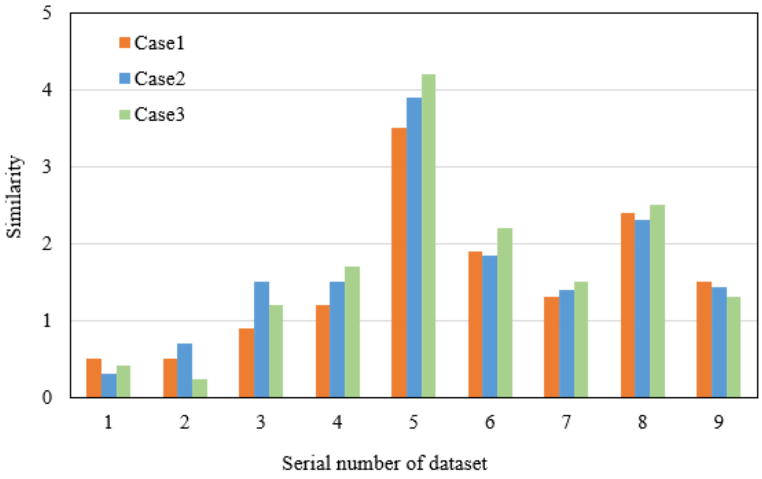


Fig. 4 Comparison of similarity of different algorithms

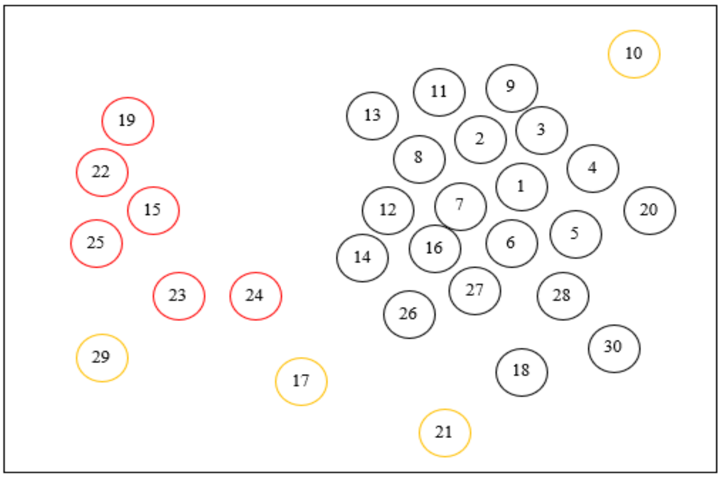


Fig. 5 Distribution of target genes in two-dimensional space

Fig. 6 shows the decision diagram of the proposed algorithm plotted based on the δ and σ values of target genes in Fig. 5. From Fig. 6, it can be clearly seen that target genes 15 and 7 have very large δ and σ values; these two target genes are density peak points and can be taken as the center points of gene clusters. Based on the decision diagram shown as Fig. 6, the center points of gene clusters can be selected manually, or the first K target genes with the greatest $\delta \times \sigma$ values can be selected as the center points of gene clusters.

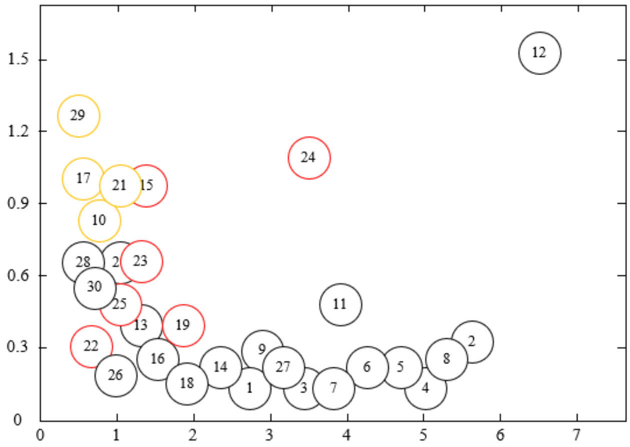
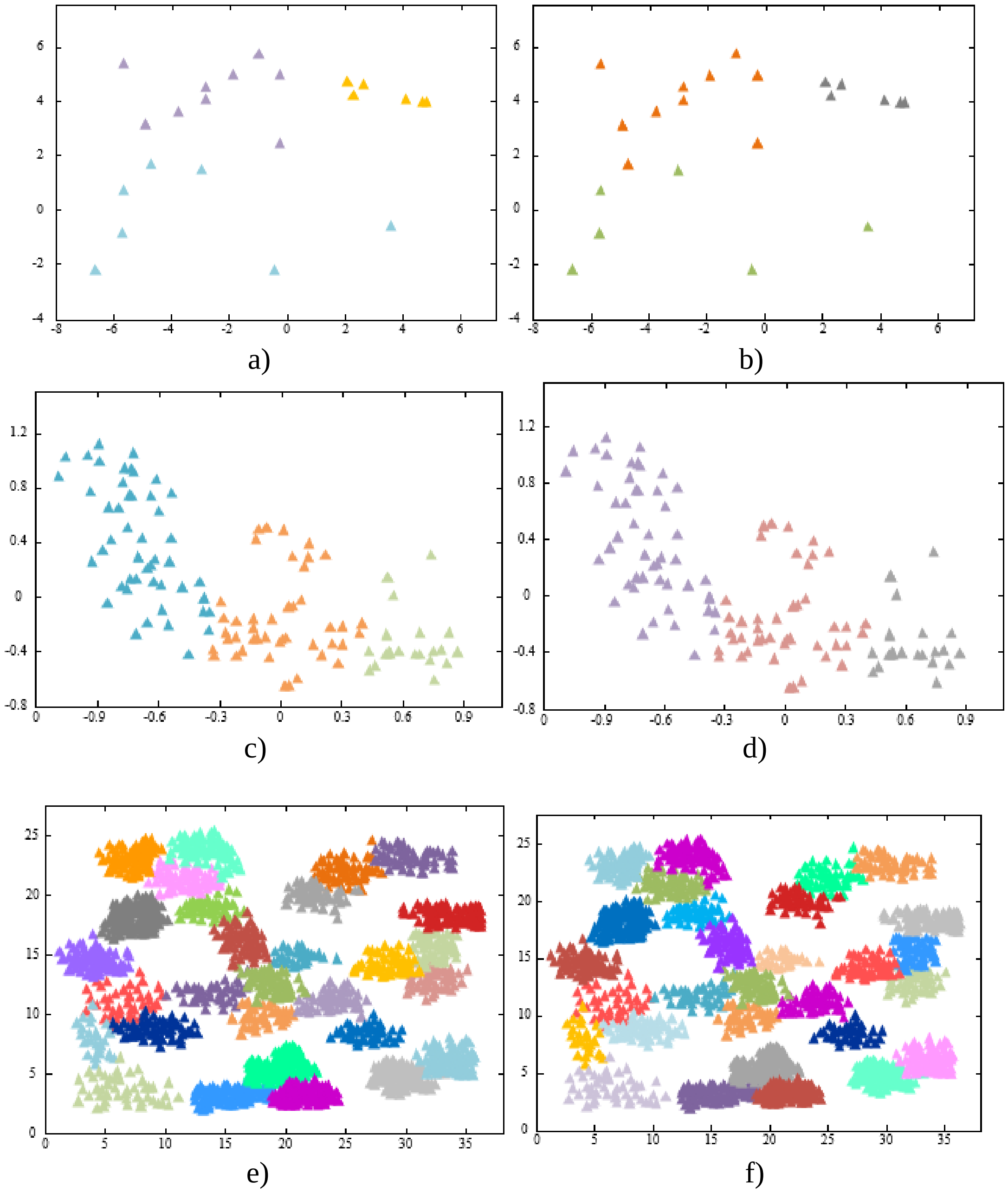


Fig. 6 Decision diagram of the center points of gene clusters

Fig. 7 visualizes the clustering results of the proposed algorithm before and after introducing the idea of similar nearest neighbor. Figs. 7a, 7c, and 7e are clustering results before introducing the idea of similar nearest neighbor under the condition of different tumor gene datasets; and Figs. 7b, 7d, and 7f are corresponding clustering results after introducing the idea of similar nearest neighbor.



a) b)

c) d)

e) f)

Fig. 7 Visualized clustering results of the algorithm before and after introducing the idea of similar nearest neighbor

According to Fig. 7, before and after introducing the idea of similar nearest neighbor, the clustering results of the proposed algorithm are basically reasonable on different datasets, but there're differences in classification accuracy. Before introducing the idea of similar nearest neighbor, the algorithm showed error clustering. However, on non-two-dimensional datasets, by comparing the clustering results of Figs. 7e and 7f, we can see that the proposed algorithm performed well before and after introducing the idea of similar nearest neighbor; before introducing the idea of similar nearest neighbor, the accuracy of the algorithm showed an obvious decline in areas with high similarity of density distribution. For crossed and repeated parts, they can be processed according to the clustering process of the proposed algorithm. Taken together, the clustering accuracy of the proposed algorithm was relatively ideal on three different types of tumor gene datasets, especially, the clustering performance was satisfactory on high-dimensional datasets.

According to above analysis, this paper adopted common artificial dataset and actual dataset for theoretical verification of the algorithm, while verifying the feasibility and effectiveness of the algorithm, this paper also applied it to actual problems to solve the difficulty of the density peak clustering algorithm in accurately selecting cluster centers in gene expression profile data, thereby proving the practical value of the proposed algorithm.

Conclusion

This paper performed cluster analysis on tumor gene data based on an improved DPC algorithm. First, the paper introduced the composition and storage form of tumor tissue samples in detail, and elaborated on the preprocessing process of gene expression profile data. Then, it carried out feature selection of tumor gene expression profile data. At last, it divided the target gene density into two parts: *K*-nearest neighbor local density, and neighborhood density. The paper fully considered the distance similarity and the density distribution similarity at the same time, and completed the improvement of the traditional DPC algorithm. In the experiment, we chose ARI and F-1 value as indexes to evaluate the clustering results of tumor gene data, and the experimental results proved that the proposed algorithm could outperform other algorithms in terms of overall performance. After that, this paper compared the similarity of different algorithms, and demonstrated that the proposed model has certain reference value in the cluster analysis of real tumor gene data. Moreover, a decision diagram of the center points of gene clusters was plotted, and clustering results of the algorithm before and after introducing the idea of similar nearest neighbor were visualized in the paper, which had verified that the accuracy of the proposed algorithm was relatively ideal, especially, its clustering performance was satisfactory on high-dimensional datasets.

References

1. Ando Y., Y. Shinozawa, Y. Iijima, B. C. Yu, M. Sone, Y. Ooi, S. I. Takahashi (2015). Tumor Necrosis Factor (TNF)-α-induced Repression of GKAP42 Protein Levels through cGMP-dependent Kinase (cGK)-Iα Causes Insulin Resistance in 3T3-L1 Adipocytes, Journal of Biological Chemistry, 290(9), 5881-5892.
2. Bathini S., S. Pakkiriswami, R. J. Ouellette, A. Ghosh, M. Packirisamy (2021). Magnetic Particle Based Liquid Biopsy Chip for Isolation of Extracellular Vesicles and Characterization by Gene Amplification, Biosensors and Bioelectronics, 194, 113585.
3. Bozic T., G. Sersa, S. K. Brezar, M. Cemazar, B. Markelc (2021). Gene Electrotransfer of Proinflammatory Chemokines CCL5 and CCL17 as a Novel Approach of Modifying Cytokine Expression Profile in the Tumor Microenvironment, Bioelectrochemistry, 140, 107795.
4. Chen G., X. Xie, S. Li (2020). Research on Complex Classification Algorithm of Breast Cancer Chip Based on SVM-RFE Gene Feature Screening, Complexity, 2020, Article ID 1342874.
5. Chen J., H. Yang, Y. Feng, Q. Shi, Z. Li, Z. Tao, X. Lu (2021). A Single Nucleotide Mutation Drastically Increases the Expression of Tumor-homing NGR-TNFα in the *E. coli* M15-pQE30 System by Improving Gene Transcription, Applied Microbiology and Biotechnology, 105(4), 1447-1460.
6. Dai L. Y., J. X. Liu, R. Zhu, X. Z. Kong, M. X.Hou, S. S. Yuan (2018). Sparse Orthogonal Nonnegative Matrix Factorization for Identifying Differentially Expressed Genes and Clustering Tumor Samples, Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1332-1337.

7. Dorosti S., S. J. Ghoushchi, E. Sobhrakhshankhah, M. Ahmadi, A. Sharifi (2020). Application of Gene Expression Programming and Sensitivity Analyses in Analyzing Effective Parameters in Gastric Cancer Tumor Size and Location, Soft Computing, 24(13), 9943-9964.
8. García-Gómez J. M., J. Gómez-Sanchis, P. Escandell-Montero, E. Fuster-Garcia, E. Soria-Olivas (2013). Sparse Manifold Clustering and Embedding to Discriminate Gene Expression Profiles of Glioblastoma and Meningioma Tumors, Computers in Biology and Medicine, 43(11), 1863-1869.
9. Hamzeh O., A. Alkhateeb, J. Zheng, S. Kandalam, L. Rueda (2020). Prediction of Tumor Location in Prostate Cancer Tissue Using a Machine Learning System on Gene Expression Data, BMC Bioinformatics, 21(2), 1-10.
10. Ji S., Y. Chen, X. Zhao, Y. Cai, X. Zhang, F. Sun, C. Liang (2021). Surface Morphology and Payload Synergistically Caused an Enhancement of the Longitudinal Relaxivity of a Mn 3 O 4/PtO x Nanocomposite for Magnetic Resonance Tumor Imaging, Biomaterials Science, 9(7), 2732-2742.
11. Jiang L. Y., D. H. Yu, X. Shi (2017). Tumor Microarray Gene Expression Data Classification Based on Weighted Extreme Learning Machine, Journal of Northeastern University (Natural Science), 38(6), 798.
12. Khalifa N. E. M., M. H. N. Taha, D. E. Ali, A. Slowik, A. E. Hassanien, (2020). Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach, IEEE Access, 8, 22874-22883.
13. Khlusov I., L. Litvinova, V. Shupletsova, O. Khaziakhmatova, E. Melashchenko, K. Yurova, Y. Sharkeev (2018). Rough Titanium Oxide Coating Prepared by Micro-arc Oxidation Causes Down-regulation of hTERT Expression, Molecular Presentation, and Cytokine Secretion in Tumor Jurkat T cells, Materials, 11(3), 360, doi: 10.3390/ma11030360.
14. Kreutz J. E., J. Wang, A. M. Sheen, A. M. Thompson, J. P. Staheli, M. R. Dyen, Q. Fengd, D. T. Chiu (2019). Self-digitization Chip for Quantitative Detection of Human Papillomavirus Gene Using Digital LAMP, Lab on a Chip, 19(6), 1035-1040.
15. Li Y., F. Fauteux, J. Zou, A. Nantel, Y. Pan (2019). Personalized Prediction of Genes with Tumor-causing Somatic Mutations Based on Multi-modal Deep Boltzmann Machine, Neurocomputing, 324, 51-62.
16. Liang X., W. Zhu, B. Liao, B. Wang, J. Yang, X. Mo, R. Li (2020). A Machine Learning Approach for Tracing Tumor Original Sites with Gene Expression Profiles, Frontiers in Bioengineering and Biotechnology, 8, 607126, doi: 10.3389/fbioe.2020.607126.
17. Masubuchi T., M. Endo, R. Iizuka, A. Iguchi, D. H. Yoon, T. Sekiguchi, H. Tadakuma, (2018). Construction of Integrated Gene Logic-chip, Nature Nanotechnology, 13(10), 933-940.
18. Meng X., Y. Yu, P. Gong, G. Jin (2021). An Integrated Droplet Digital PCR Gene Chip for Absolute Quantification of Nucleic Acid, Microfluidics and Nanofluidics, 25(7), 1-9.
19. Otuboah F. Y., Z. Jihong, Z. Tianyun, C. Cheng (2019). Design of a Reduced Objective Lens Fluorescence dPCR Gene Chip Detection System with High-throughput and Large Field of View, Optik, 179, 1071-1083.
20. Otuboah F. Y., J. Zheng, C. Chen, Z. Wang, X. Wan, L. Sun (2020). High-throughput and Uniform Large Field-of-view Multichannel Fluorescence Microscopy with Super-thin Dichroism for a dPCR Gene Chip, Applied Optics, 59(34), 10768-10776.
21. Pandey R., O. Teig-Sussholz, S. Schuster, A. Avni, Y. Shacham-Diamand (2018). Integrated Electrochemical Chip-on-plant Functional Sensor for Monitoring Gene Expression under Stress, Biosensors and Bioelectronics, 117, 493-500.

22. Qaddoum K. (2018). Gene Selection Approach Utilizing Data Clustering Based Technique Optimization for Tumor Classification, Proceedings of the 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 3, 128-135.
23. Qian Y., L. Zhang, M. Cai, H. Li, H. Xu, H. Yang, W. Lu (2019). The Prostate Cancer Risk Variant rs55958994 regulates Multiple Gene Expression through Extreme Long-range Chromatin Interaction to Control Tumor Progression, Science Advances, 5(7), eaaw6710.
24. Sarafidis M., A. Zaravinos, D. Iliopoulou, D. Koutsouris, G. I. Lambrou (2019). Regressions of Clustered Gene Expression Data Manifest Tumor-specific Genes in Urinary Bladder Cancer, Proceedings of the 19th International Conference on Bioinformatics and Bioengineering (BIBE), 127-131.
25. Tsai W. K., C. I. Wang, C. H. Liao, C. N. Yao, T. J. Kuo, M. H. Liu, Y. H. Chan (2019). Molecular Design of Near-infrared Fluorescent Pdots for Tumor Targeting: Aggregation-induced Emission versus Anti-aggregation-caused Quenching, Chemical Science, 10(1), 198-207.
26. Wang A., N. An, G. Chen, L. Liu, G. Alterovitz (2018). Subtype Dependent Biomarker Identification and Tumor Classification from Gene Expression Profiles, Knowledge-based Systems, 146, 104-117.
27. Wang X., J. Liu, Y. Cheng, A. Liu, E. Chen (2018). Dual Hypergraph Regularized PCA for Biclustering of Tumor Gene Expression Data, IEEE Transactions on Knowledge and Data Engineering, 31(12), 2292-2303.
28. Yang C., Y. Deng, H. Ren, R. Wang, X. Li (2020). A Multi-channel Polymerase Chain Reaction Lab-on-a-chip and Its Application in Spaceflight Experiment for the Study of Gene Mutation, Acta Astronautica, 166, 590-598.
29. Yu Z., H. Chen, J. You, H. S. Wong, J. Liu, L. Li, G. Han (2014). Double Selection Based Semi-supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(4), 727-740.
30. Yuan D., J. Kong, X. Fang, Q. Chen (2019). A Graphene Oxide-based Paper Chip Integrated with the Hybridization Chain Reaction for Peanut and Soybean Allergen Gene Detection, Talanta, 196, 64-70.
31. Zaritski A., H. Castillo-Ecija, M. Kumarasamy, E. Peled, R. Sverdlov Arzi, A. M. Carcaboso, A. Sosnik (2019). Selective Accumulation of Galactomannan Amphiphilic Nanomaterials in Pediatric Solid Tumor Xenografts Correlates with GLUT1 Gene Expression, ACS Applied Materials and Interfaces, 11(42), 38483-38496.

Wei Wang, M.Sc.
E-mail: qhdwang2020@163.com

Wei Wang was graduated from Hebei Normal University, China, and now serves as a lecturer at Computer Department of Qinhuangdao Radio and Television University, China. His research direction is computer application.

Bo Gao, M.Sc.
E-mail: gabosky@163.com

Bo Gao was graduated from Beijing University of Aeronautics and Astronautics, China, and now serves as lecturer at Computer Department of Qinhuangdao Radio and Television University, China. His research direction is computer software.