


A Unified Multi-View Clustering Method Based on Non-Negative Matrix Factorization for Cancer Subtyping

Zhanpeng Huang, Guangdong University of Technology, China*

 <https://orcid.org/0000-0003-1888-7490>

Jiekang Wu, School of Automation, Guangdong University of Technology, China

Jinlin Wang, Guangzhou Medical University, China

Yu Lin, Southern Medical University, China

Xiaohua Chen, Southern Medical University, China

ABSTRACT

Non-negative matrix factorization (NMF) has gained sustaining attention due to its compact learning ability. Cancer subtyping is important for cancer prognosis analysis and clinical precision treatment. Integrating multi-omics data for cancer subtyping is beneficial to uncover the characteristics of cancer at the system-level. A unified multi-view clustering method was developed via adaptive graph and sparsity regularized non-negative matrix factorization (multi-GSNMF) for cancer subtyping. The local geometrical structures of each omics data were incorporated into the procedures of common consensus matrix learning, and the sparsity constraints were used to reduce the effect of noise and outliers in bioinformatics datasets. The performances of multi-GSNMF were evaluated on ten cancer datasets. Compared with 10 state-of-the-art multi-view clustering algorithms, multi-GSNMF performed better by providing significantly different survival in 7 out of 10 cancer datasets, the highest among all the compared methods.

KEYWORDS

Cancer Subtyping, Graph Regularized, Multi-View Clustering, Non-Negative Matrix Factorization, Sparsity Regularized

INTRODUCTION

Due to the increasing number of new cancer cases and deaths, even with the rapid development of medical technology, cancer still seriously threatens human health and is an important cause of human death. The latest estimates for cancer from the International Agency for Research on Cancer (IARC, 2021) show 19.3 million new cases of cancer worldwide and 10 million cancer deaths in 2020. Cancer is expected to surpass cardiovascular disease as the main cause of premature death in most countries in this century. The rapid development of high-throughput technologies such as deep sequencing

DOI: 10.4018/IJDWM.319956

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

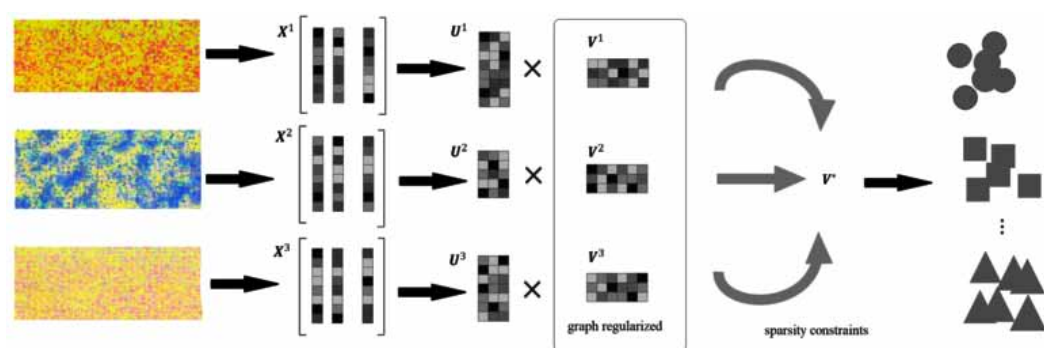
has enabled the discovery of mass amounts of biological information, which is conducive to better characterizing human diseases and facilitating personalized treatments. In oncology, analysis based on high-throughput biological data sets has discovered new cancer subtypes, which have been used for cancer treatment decisions (Parker et al., 2009; Prasad et al., 2016).

Machine learning technology is widely used in the analysis of bioinformatics data, which can support decision-making and treatment planning for the doctors (Amin et al., 2021; Kumar-Sinha & Chinnaiyan, 2018; Rajinikanth & Kadry, 2021). In order to improve cancer diagnosis and treatment, genomic and other molecular profiles of tumor biopsies have been analyzed for precision tumor therapy. By incorporating gene network interaction, a novel coclustering algorithm has been proposed for identifying cancer subtypes (Liu et al., 2014). However, the role of the human genome is complex and chaotic, and it can regulate biological processes at different levels. The human genome could be revealed by integrating various genomics, such as gene expression, copy number variation, and DNA methylation (Huang et al., 2017). Modern genomic and clinical research urgently needs integrated machine learning models of multiomics data to better utilize large amounts of heterogeneous information to deeply understand biological systems. Multiomics data can obtain information from different perspectives and levels, which is conducive to understanding complex biological systems (Li et al., 2016). The integration and clustering of multiomic data are some of the research hotspots of machine learning in the field of bioinformatics.

To take advantage of local geometrical structures and global structures of the bioinformatics data, a novel multiview clustering method based on nonnegative matrix factorization (NMF) is proposed for cancer subtyping. The local geometrical structures of each omics data set were encoded by generating a nearest neighbor graph. The global structures of a multiomics data set were captured by the sparsity regularized constraints. Then, the unified objective function was used by incorporating local geometrical structures of each omics data set and sparsity regularized common consensus matrix into the NMF-based framework. The novel multiview NMF-based method can obtain the common consensus representation of a multiomics data set, while the sparsity constraints are integrated to handle the noise and outliers in bioinformatics data. Figure 1 illustrates the framework of the unified multiview clustering method. The multiview NMF with graph-regularized and sparsity constraints was integrated to form a unified framework. The final clustering results were gained by spectral clustering. The main contributions are as follows:

1. A unified framework for cancer subtyping by considering the feature of a cancer data set was proposed, which will be useful to identify cancer subtyping in precision medicine that would otherwise be obscured by noise and outliers in bioinformatics.

Figure 1. Framework of the Proposed Algorithm



2. The local geometrical structures and sparsity constraints are incorporated into the multiview clustering process to form a unified objective function for cancer subtyping based on nonnegative matrix factorization.
3. By incorporating the local geometrical structures of each omics data set and the sparsity constraints on a common consensus matrix into the clustering process, Multi-GSNMF provides a unified model and a novel solution to fuse multiview data for clustering.

RELATED WORK

The advent of high-throughput sequencing technologies has allowed the development of numerous multiomics clustering methods. At the beginning, some multiomics methods were developed by extending basic clustering algorithms in the multiomics clustering field, for example, k-means (Bickel & Scheffer, 2004). Multiomics clustering has been divided into early integration, late integration, and intermediate integration. Early integration methods are to directly concatenate multiomics data to form a matrix with all features. For example, LRAcluster uses a low-rank approximation by probabilistic model to find the shared principal subspace across multiomics data sets (Wu et al., 2015). The advantage is that it includes biological knowledge. The iCluster is also an early integration approach, which assumes a regularized joint latent variable and projects the data to a lower dimension by probabilistic modeling (Shen et al., 2009, 2012). As an extension of iCluster, iClusterBayes was proposed to concatenate multiomics data of different types by using a Bayesian latent variable for cancer subtypes (Mo et al., 2018). However, the early integration methods were unable to handle the different distributions of different omics data, which greatly affects the clustering effect of the correlation algorithm.

The late integration methods use existing single-view clustering on single omics data. Then, the different clustering results are integrated together. Cluster-of-cluster assignments (COCA, Hoadley et al., 2014) and data integration and cancer subtyping by perturbation clustering (PINS, Nguyen et al., 2017) fall into this category. The advantage of late integration is that different clustering algorithms can be chosen according to the characteristics of single-omics data. PINS first computes the binary matrix to gain a sample relationship matrix. Then, tests whether the obtained clusters can be split into smaller clusters. It integrates clusters by examining the connectivity matrices of different omics. The late integration methods divide multiview clustering into two steps, which can reduce the computational complexity of algorithms. However, it is difficult to make full use of the consistency and difference between each omics data set.

Intermediate integration methods try to construct a framework that integrates the information of all omics data, and most multiomics clustering methods fall into this category. Similarity-based methods and dimensionality reduction-based methods are intermediate integration methods. Similarity-based methods include spectral (Chikhi, 2016), similarity network fusion (SNF, Wang et al., 2014), and neighborhood-based multiomics clustering (NEMO, Rappoport & Shamir, 2019). SNF builds a similarity matrix between samples for every omics and fuses all the matrices into one consensus matrix. NEMO performs fusion according to the distance similarity between samples to obtain a consensus matrix between multiomics data.

Dimensionality reduction-based methods transform data from a high-dimensional space into a low-dimensional space. In bioinformatics applications, the omics data sets are often high dimensionality. The canonical correlation analysis method (CCA, Chaudhuri et al., 2009), partial least squares (PLSs) method (Lê et al., 2009), multiple kernel learning method (rMKL-LPP, Speicher & Pfeifer, 2015), and NMF method are the most widely used dimension reduction methods. However, CCA can only handle data from two views. MCCA uses multiset canonical correlation analysis by maximizing the sum of the pairwise correlation between projections, which extends CCA to more than two views (MCCA, Witten & Tibshirani, 2009).

rMKL-LPP efficiently captured the similarity between different samples by mapping them to a high dimension. Multiple kernels were used to learn the information of each omics, and then they were linearly combined. NMF is a common dimensionality reduction method. An NMF-based multiview clustering method was proposed with consistency constraint, which integrates each view's representation toward a common consensus, and was formulated to handle multiomics data clustering (MultiNMF, Liu et al., 2013). A unified model is used in dimensionality reduction methods, which makes full use of information within and between omics. At the same time, the computational complexity of the algorithms is also reasonable by using dimension reduction.

The NMF has been introduced as a dimension reduction method (Lee & Seung, 1999). In the real world, many data are nonnegative, such as gene expression data and image pixel data. Thus, NMF assumes that data have a low-dimensional nonnegative representation, which has been widely used in pattern recognition (Wen et al., 2018; Sun et al., 2016), bioinformatics (Want et al., 2022), and computer vision (Li et al., 2021). Based on the MultiNMF methods proposed by Liu et al. (2013), many manifold learning and pairwise measurement technologies are used on NMF-based multiview clustering methods (Wang et al., 2019; Liang et al., 2020). However, with the increase of data dimensions, it becomes increasingly difficult to find meaningful clustering results (Janeja et al., 2020). For the clustering of high-dimensional data sets, sparsity constraints are usually used to identify the global structures of data sets (Huang & Wu, 2022). The distinguishing feature of bioinformatics data is the small number of samples relative to the large number of features. In the biological process, different levels of omics data have different statistical properties and distribution structures. How to make full use of the local features of each omics data set to construct a global consensus matrix for all omics data is a challenging research problem.

A Unified Multiview Clustering Framework via NMF

Multiview Nonnegative Matrix Factorization

Given a nonnegative data matrix, $X = [x_1, x_2, \dots, x_n] \in R^{M \times N}$, where each column $x_i \in R^m$ ($i = 1, 2, \dots, n$) represents a data point where M is the dimension of the feature and N refers to the number of data points, NMF aims to find two nonnegative matrix factors $U = [U_{i,k}] \in R^{M \times k}$ and $V = [V_{j,k}] \in R^{N \times k}$, whose product is a good approximate to X ; $X \approx UV^T$, and k is the designed dimensionality, $k \ll \min(M, N)$. On account of learning compact representation, U denotes the basis matrix and V can be interpreted as a coefficient matrix.

Frobenius norm (Paatero & Tapper, 1994) and Kullback–Leibler divergence (Lee & Seung, 2000) are the two commonly used cost functions for quantifying the quality of the approximation of X . The Frobenius norm cost function is used in this paper, which is defined as:

$$\begin{aligned} \min_{U, V} \|X - UV^T\|_F^2 \\ \text{s.t. } U \geq 0, V \geq 0 \end{aligned} \quad (1)$$

Multiview clustering via joint NMF was proposed by searching for a factorization that gives compatible clustering solutions across multiple views to integrate information from multiple views in the unsupervised setting (Liu et al., 2013). In single-view NMF, coefficient matrix V can be regarded as a low-rank representation of data points in terms of the new basis U . The loss function $D(V^v, V^*) = V^v - V^{*2}_F$ is used as a measure of the disagreement between coefficient matrixes V^v and V^* , which denote the coefficient matrix of the v -th view and the consensus matrix, respectively. The joint minimization problem for multiview NMF is as follows:

$$\sum_{v=1}^{n_v} X^v - U^v \left(V^v \right)_F^{T2} + \sum_{v=1}^{n_v} \lambda_v V^v - V^{*2}_F$$

$$s.t. \ U^v > 0, V^v > 0, V^* > 0, U^{v,*}_{*,k1} = 1, \lambda_v > 0, v = 1, \dots, n_v \quad (2)$$

where λ_v is the parameter to adjust the relative weight among different views while adjusting the standard NMF reconstruction error and disagreement term $D(V^v, V^*)$ and n_v is the number of views. A diagonal matrix Q^v is introduced to simplify the computation to remove the equality constraint on U^v :

$$Q^v = \text{Diag} \left(\sum_{i=1}^M U^v_{i,1}, \sum_{i=1}^M U^v_{i,2}, \dots, \sum_{i=1}^M U^v_{i,K} \right) \quad (3)$$

where $\text{Diag}(\ast)$ denotes a diagonal matrix with nonzero elements equal to the values in the parenthesis sequentially. Then the equality constraint on U^v can be removed and the objective function for multiview NMF can be defined as follows:

$$\sum_{v=1}^{n_v} X^v - U^v \left(V^v \right)_F^{T2} + \sum_{v=1}^{n_v} \lambda_v V^v Q^v - V^{*2}_F$$

$$s.t. \ U^v > 0, V^v > 0, V^* > 0, \lambda_v > 0, v = 1, \dots, n_v \quad (4)$$

The coefficient matrix is learned from factorization of each view and is regularized toward a common consensus matrix. The consensus matrix is thought to reflect the underlying clustering structure shared by different views.

Framework Overview of Multi-GSNMF

Let X^1, X^2, \dots, X^{n_v} be the data of all views, $v = 1, 2, \dots, n_v$, and k be the desired reduced dimension. The numbers of samples are the same for all the views, while different numbers of the features are allowed. The goal of Multi-GSNMF is to form a consensus matrix among all views by integrating the relationship between sample data from different views. The same as multiview NMF, the inconsistency penalty function is incorporated into the NMF framework. In order to make full use of the intrinsic geometric information between samples, the local geometric structures are used for each view NMF, which is inspired by graph-regularized nonnegative matrix factorization (GNMF, Cai et al., 2011). GNMF constructs a nearest neighbor graph to model the manifold structures and provides a principled way for incorporating the geometrical structures into the model. By using the Euclidean distance, the objective function of GNMF is as follows:

$$X - UV^{T2} + \tau \text{Tr} \left(V^T L V \right) \quad (5)$$

where τ adjust the smoothness of the presentation. In the equation, $\text{Tr}(\cdot)$ is the trace of a matrix and \mathbf{D} is a diagonal matrix, which is defined as $D_{ij} = \sum_i W_{ij}$ and denotes the degree matrix $L = D - W$.

Therefore, by incorporating the local geometric structures, the multiview NMF can be formulated as:

$$\sum_{v=1}^{n_v} \left(X^v - U^v \left(V^v \right)_F^{T2} + a_v \text{Tr} \left(\left(V^v \right)^T L V^v \right) \right) + \eta \sum_{v=1}^{n_v} a_v V^v Q^v - V^{*2}_F$$

(6)

where a_v is the parameter to adjust the weights between NMF and local geometric structures for each view. Considering that the weights of each omics data set in the object function should be the same, the a_v is used in the consistency loss punishment function of each omics data set. The η is used to balance the effect of Multi-GSNMF and the consistency loss punishment function, and the final weights of each omics for consistency loss punishment function is $\eta^* a_v$.

Many studies have shown that sparse constraints can achieve better robustness and improve the clustering performances (Hoyer, 2004; Elhamifar & Vidal 2013). Similar to other high-dimensional data, there are noises and outliers in bioinformatics data. However, in the objective function of NMF, the error of each sample point is the square residual, which leads to outliers with large errors that will greatly affect the objective function (Liu et al., 2018). $\ell_{2,1}$ -norm is robust to noise and outliers; therefore, the spare constraints by $\ell_{2,1}$ regularization is added to the consensus matrix V^* , which is widely used in many applications (Li et al., 2021). Therefore, the optimization goal can be attained as follows by integrating the above parts into a unified objective function:

$$\mathcal{O} = \min \sum_{v=1}^{n_v} \left(X^v - U^v \left(V^v \right)_F^{T2} + a_v \text{Tr} \left(\left(V^v \right)^T L V^v \right) \right) + \eta \sum_{v=1}^{n_v} a_v V^v Q^v - V^{*2}_F + \lambda V^*_{2,1}$$

(7)

where λ is the parameter to adjust the weight of sparse constraints.

Optimization Algorithm for Multi-GSNMF

Fixing V^* and Minimizing \mathcal{O} Over U^v and V^v

Each view is independent when V^* is given. The calculation of U^v does not rely on $U^{v'}$ or $V^{v'}$, $v' \neq v$. Therefore, \mathbf{X} , \mathbf{U} , \mathbf{V} , and \mathbf{Q} are used to denote X^v , U^v , V^v , and Q^v for simplicity in this subsection. Then, the objective function \mathcal{O} in Equation (7) can be minimized as follows:

$$\mathcal{O}_1 = X - UV^{T2}_F + a_v \text{Tr} \left(V^T L V \right) + \eta^* a_v V Q - V^{*2}_F \quad s.t. \quad U^v > 0, V^v > 0$$

(8)

The objective function \mathcal{O}_1 in Equation (8) is not convex with U and V . As Paatero & Tapper (1994) suggested, finding the global minima is difficult. The following procedures updates the

values of appointed variables by fixing some variables sequentially and alliteratively to achieve the local minima.

Fixing V^{} and V^v and Updating U^v*

Let φ_{ik} be the Lagrangian multiplier for constraint $u_{ik} \geq 0$ and $\psi = [\varphi_{ik}]$. \mathcal{L}_1 is the Lagrangian $\mathcal{L} = \mathcal{O}_1 + Tr(\psi U)$, where $Tr(\cdot)$ is the trace function. When it is neglecting the constants, the Lagrangian function is written as follows:

$$\mathcal{L}_1 = Tr(UV^T V U^T) - 2Tr(XV U^T) + \eta^* a_v R + Tr(\psi U) \quad (9)$$

where $R = Tr[VQ Q^T V^T - 2VQ(V^*)^T]$. And the derivative of R with respect to U is:

$$P_{i,k} = \frac{\partial R}{\partial U_{i,k}} = 2 \left(\sum_{m=1}^M U_{m,k} \sum_{j=1}^N U_{j,k}^2 - \sum_{j=1}^N V_{j,k} V_{j,k}^* \right) \quad (10)$$

The partial derivatives of \mathcal{L}_1 with respect to U is:

$$\frac{\partial \mathcal{L}_1}{\partial U} = 2UV^T V - 2XV + \eta^* a_v P + \psi \quad (11)$$

Using the Karush–Kuhn–Tucker (KKT) (Boyd & Vandenberghe, 2004) conditions $\varphi_{ik} u_{ik} = 0$, the following updating rule is obtained:

$$u_{i,k} \leftarrow u_{i,k} \frac{(XV)_{i,k} + \eta^* a_v \sum_{j=1}^N V_{j,k} V_{j,k}^*}{(UV^T V)_{i,k} + \eta^* a_v \sum_{m=1}^M U_{m,k} \sum_{j=1}^N V_{j,k}^2} \quad (12)$$

Fixing V^{} and U^v and Updating V^v*

When computing V^v , the column vectors, U^v and V^v , are normalized first using Q^v as Equation (4) defined:

$$U \leftarrow UQ^{-1}, V \leftarrow VQ \quad (13)$$

Let ϕ_{jk} be the Lagrangian multiplier for constraint $v_{jk} \geq 0$ and $\Phi = [\phi_{jk}]$. The Lagrangian function is written as follows:

$$\mathcal{L}_2 = Tr(UV^T V U^T - 2XV U^T) + a_v Tr(V^T L V) + \eta^* a_v Tr(VV^T - 2V(V^*)^T) + Tr(\Phi V) \quad (14)$$

The partial derivatives of \mathcal{L}_2 with respect to V are:

$$\frac{\partial \mathcal{L}_2}{\partial V} = 2VU^T U - 2X^T U + 2a_v L V + 2\eta^* a_v (V - V^*) + \Phi \quad (15)$$

By using the KKT condition, $\phi_{jk} v_{jk} = 0$, the solution results in the updating rule as follows:

$$v_{j,k} \leftarrow v_{j,k} \frac{(XU)_{j,k} + \eta^* a_v V_{j,k}^* + a_v (WV)_{j,k}}{(VU^T U)_{j,k} + \eta^* a_v V + a_v (DV)_{j,k}} \quad (16)$$

where \mathbf{W} is a weight matrix and \mathbf{D} is a diagonal matrix. The entries of \mathbf{D} are the column sums of the weight matrix \mathbf{W} .

Fixing U^v and V^v and Updating V^*

When U^v and V^v are fixed, minimizing the objective equation \mathcal{O} is equivalent to solving the minimum values of the equation:

$$\mathcal{O}_2 = \eta \sum_{v=1}^{n_v} a_v V^v Q^v - V_F^{*2} + \lambda V_{2,1}^* \quad (17)$$

Let Σ^v be the Lagrangian multiplier for nonnegative constraint $V^* \geq 0$. The Lagrangian function of Equation (17) can be rewritten as follows:

$$\mathcal{L}_3 = \eta \sum_{v=1}^{n_v} a_v V^v Q^v - V_F^{*2} + \lambda V_{2,1}^* + Tr(\Sigma^v V^*) \quad (18)$$

The partial derivatives of \mathcal{L}_3 with respect to V^* were calculated, and the KKT conditions $\varsigma_{ij}^v v_{ij}^* = 0$ were used to obtain the equation:

$$2\eta \sum_{v=1}^{n_v} a_v (V^* - V^v) v_{ij}^* + \frac{\lambda v_{ij}^*}{\sqrt{\sum_{k=1}^K (v_{ij}^*)^2}} v_{ij}^* = 0 \quad (19)$$

Finally, the updating rule for v_{ij}^* is:

$$v_{ij}^* \leftarrow v_{ij}^* \frac{2\eta \sum_{v=1}^{n_v} a_v V^v}{2\eta \sum_{v=1}^{n_v} a_v V^v + \frac{\lambda v_{ij}^*}{\sqrt{\sum_{k=1}^K (v_{ij}^*)^2}}} \quad (20)$$

The steps of updating were repeated until convergence or the maximal iterations. The convergence condition is defined as the relative change ratio of objective function: $\mathcal{O}_t - \mathcal{O}_{t-1} / \mathcal{O}_t \leq \delta$, which is set to 10^{-3} in all experiments. The maximal iterations μ_{\max} is simultaneously set to 50.

The U^v , V^v , and V^* for each omics data set need to be initialized before the loop iteration. To get the manifold structures for each omics data set, the U^v and V^v are calculated by the graph-regularized nonnegative matrix factorization (Cai et al., 2011). And the heat kernel weighting is chosen for constructing the weight matrix. After the U^v and V^v have been initialized, the initial matrix of V^* can be gained by the mean of all the coefficient matrix for each view.

Experiment

We present the procedures of the proposed Multi-GSNMF approach in Algorithm 1 (see Table 1). Multi-GSNMF was implemented on the Matlab 2020a platform. To demonstrate the ability of Multi-GSNMF on multiview clustering, the performance of Multi-GSNMF on 10 multiomics cancer data sets has been evaluated.

Parameter Estimation

The three parameters were k , a_v , η , and λ . k is the clusters number of the multiomics data. η and λ represent the inconsistency punishment function and the sparsity regularization constants for

Table 1. Algorithm 1: Procedures of the Multi-GSNMF for Cancer Subtyping

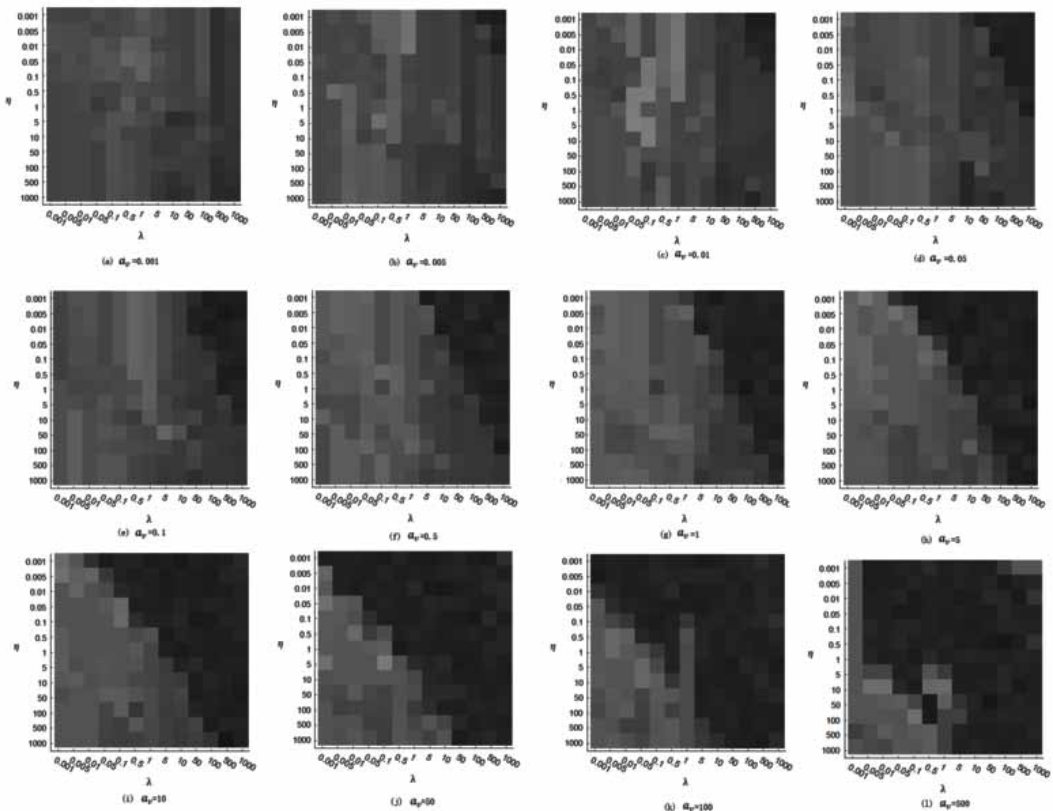
<p>Require: Multiomics data set $\left\{X^{(v)}\right\}_{v=1}^{n_v}$, parameters k, a_v, η, and λ.</p> <p>Ensure: Consensus Matrix V^*.</p> <ol style="list-style-type: none"> 1. Normalizing each view $X^{(v)}$, such that $X^{(v)}_1 = 1$ 2. Initialization: U^v, V^v, and V^* ($1 \leq v \leq n_v$), $\delta = 10^{-3}$, $\mu_{\max} = 50$. 3. Repeat 4. For $v = 1$ to n_v 5. Repeat 6. Update U^v by Equation (12) 7. Normalizing U^v and V^v by Equation (13) 8. Update V^v by Equation (16) 9. Until Equation (8) convergence. 10. End for 11. Update V^* by Equation (20) 12. Until Equation (7) convergence or reach maximal iterations.
--

the Multi-GSNMF method, respectively, and a_v is the parameter used to adjust the weights between different omics data sets.

The proper parameters need to be chosen for Multi-GSNMF before being evaluated on the real-world cancer data sets. As cancer subtyping is continuously optimized, there is no standard subtyping for biological data sets. A small number of samples versus a large number of features and the low-dimensional representation are the two prominent characteristics of biological data sets. The 3 Sources Dataset (April 2009) has similar characteristics with biological data sets, so it is chosen as a training set for parameters tuning. The 3 Sources Dataset is a data set of news articles collected from the *British Broadcasting Corporation (BBC)*, *Reuters*, and *The Guardian*. There are 169 articles across six topic categories from all three sources. Similar to bioinformatics data sets, each article has thousands of features specific to the news sources. Accuracy (AC, Xu et al., 2003) was selected as the criterion to evaluate the performances of Multi-GSNMF with different parameters. AC calculates the consistency between the clustering results of Multi-GSNMF and the actual classification.

In order to choose the proper initial parameters for the proposed algorithm, a wide ranges of grids $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000\}$ have been searched for the initial a_v , η , and λ . The 3 Sources Dataset was used as the training set for parameter tuning. Figure 2 shows the performance of Multi-GSNMF on the 3 Sources Dataset, some results are not displayed due to the size of the figure. η and λ are the trade-off parameters. Each small figure represents the AC value of a_v by alternately varying the parameters η and λ . The horizontal axis represents the

Figure 2. ACC of Multi-GSNMF With Different a_v , η , and λ in a Wide Range of Grids



change of η , and the vertical axis represents λ . The higher the value of AC, the brighter the picture. As we can see in Figure 2(c), Multi-GSNMF gains better performance for a block of higher AC values when $a_v = 0.01$ and a striped highlight area when $\lambda = 1$. And the value $\eta = 0.05$ in the middle of the striped highlight area is chosen. When $a_v = 0.01$, $\eta = 0.05$, and $\lambda = 1$, Multi-GSNMF achieves a better AC value on the training set, and the AC values around it do not decrease seriously. So, the initial parameters chosen for the real cancer data set were $a_v = 0.01$, $\eta = 0.05$, and $\lambda = 1$.

Computational Complexity and Convergence Analysis

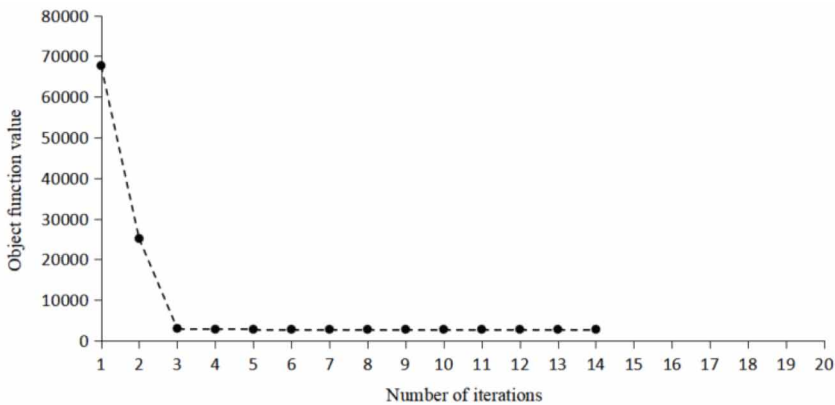
For each single view, Steps from 6 to 8 are repeated until convergence in Algorithm 1 (see Table 1). The computational complexity of each iteration for each view is $O(MNk)$, where M is the dimension of the feature and N refers to the number of data points, k is the number of clusters. There are t_{in} loops and n_v views, so the computational complexity of Steps 4 to 10 is $O(t_{in} n_v MNk)$. Multi-GSNMF also needs $O(n_v Nk)$ to compute the consensus matrix according to Equation (20). Assume there are t_{out} loops for Steps 3 to 12 in Algorithm 1, the overall computational complexity of Multi-GSNMF is $O(t_{out}(t_{in} n_v MNk + n_v Nk))$.

In our method, the U^v , V^v , and V^* update in each iteration. To show the convergence property of Multi-GSNMF, the objective function values of the 3 Sources Dataset in each iteration are shown in Figure 3. It shows that the objective function values drop rapidly and converge quickly.

Evaluation and Comparison on the Cancer Data Set

The Multi-GSNMF algorithm was evaluated on 10 multiomics cancer data sets. The bioinformatics data are publicly available at The Cancer Genome Atlas (TCGA), which provides a massive amount of bioinformatics data (Weinstein et al., 2013). The Acute Myeloid Leukemia (AML), Breast invasive carcinoma (BIC), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Kidney renal clear cell carcinoma (KIRC), Liver hepatocellular carcinoma (LIHC), Lung squamous cell carcinoma (LUSC), Skin cutaneous melanoma (SKCM), Ovarian serous cystadenocarcinoma (OV), and Sarcoma (SARC) are used for algorithm comparison, which are provided by Rappoport and Shamir (2018). The number of samples in each cancer data set range from 170 for AML to 621 for BIC, which all contain three omics: gene expression, DNA methylation, and miRNA expression.

Figure 3. The Convergence Curves of the Objective Values With Respect to Iteration Time



Several typical multiomics clustering methods were chosen to compare the performances on 10 cancer data sets with Multi-GSNMF. All comparison algorithms have been evaluated based on the platform developed by Rappoport and Shamir (2018), while NEMO used the source code provided by the authors. The parameters suggested by the authors were used. Each method determines the number of clusters for each data set. However, to choose the number of clusters used by MCCA and LRAcluster, Rappoport and Shamir mistakenly used the minimal silhouette score. The maximal silhouette score suggested by the authors led to worse performances for both methods, so we used the version provided by Rappoport and Shamir (2018).

The actual clustering numbers are unknown for each real-world cancer data set. Different survival rates between the subtyping for each cancer data set has been measured using the log-rank test. We assumed that if patients between different subtypes have significantly different survival rates, they were different in determining the biological significance of the clustering numbers. Therefore, the p value derived for the log-rank test is used as a metric for the choice of the clustering numbers. The proposed algorithm has been run with 2–10 clusters. The minimum of the p values was chosen as the suggested cluster's number, which are presented in Table 2. The final performances of Multi-GSNMF are based on the chosen cluster's number.

The numbers of enriched clinical labels and the clustering with significantly different survival rates are used as the criteria to evaluate the performance of each clustering method (Rappoport and Shamir (2018)). The survival rate differences between patients of different subtyping are computed by a log-rank test. A clustering is considered as biologically significant if the patients of different subtyping have significant survival rate differences ($p \leq .05$). Meanwhile, enrichment of clinical labels for the patient with different subtyping are computed. The χ^2 test of independence is calculated for enrichment of discrete variables, and Kruskal–Wallis test for enrichment of numeric variables.

Figure 4 shows the performances of Multi-GSNMF and the other 10 compared methods on the 10 real-world cancer data sets. The $-\log_{10}$ values of survival rate difference by log-rank test between subtyping are shown on the x -axis, while the y -axis corresponds to the numbers of enriched clinical labels. The dotted line parallel to the y -axis is the dividing line for whether the survival rate difference is significant or not. Values on the x -axis greater than the dividing line indicate that the cancer subtyping of the clustering algorithm has significant survival rate differences. For the COAD and LUSC data sets, all tested algorithms failed to find clustering with significant survival rate differences. Multi-GSNMF found clustering with significant survival rate differences in 7 out of 10 cancer types, except COAD, LUSC, and SKCM. Multi-GSNMF outperforms the compared methods at finding clustering with significant survival rate differences, while all the other methods found clustering with significantly different survival rates in at most six cancer data sets.

On the OV data set, the subtyping of Multi-GSNMF was found to have significantly different survival rates, and at the same time, one enriched clinical label. Multi-GSNMF gained the largest number of enriched clinical labels on the KIRC data set and was also found to have the smallest p value of different survival rates on the GBM and OV cancer data sets. However, the proposed algorithm fails to find an enriched clinical label on SKCM.

Table 2. Clustering Numbers Chosen by the Multi-GSNMF on 10 Cancer Data Sets

Data sets	Acute Myeloid Leukemia	Breast invasive carcinoma	Colon adenocarcinoma	Glioblastoma multiforme	Kidney renal clear cell carcinoma	Liver hepatocellular carcinoma	Lung squamous cell carcinoma	Skin cutaneous melanoma	Ovarian serous cystadenocarcinoma	Sarcoma
Number of clusters	3	3	5	7	5	10	3	5	4	10

Figure 4. Performances of all the Testing Methods on 10 Multiomics Cancer Data Sets

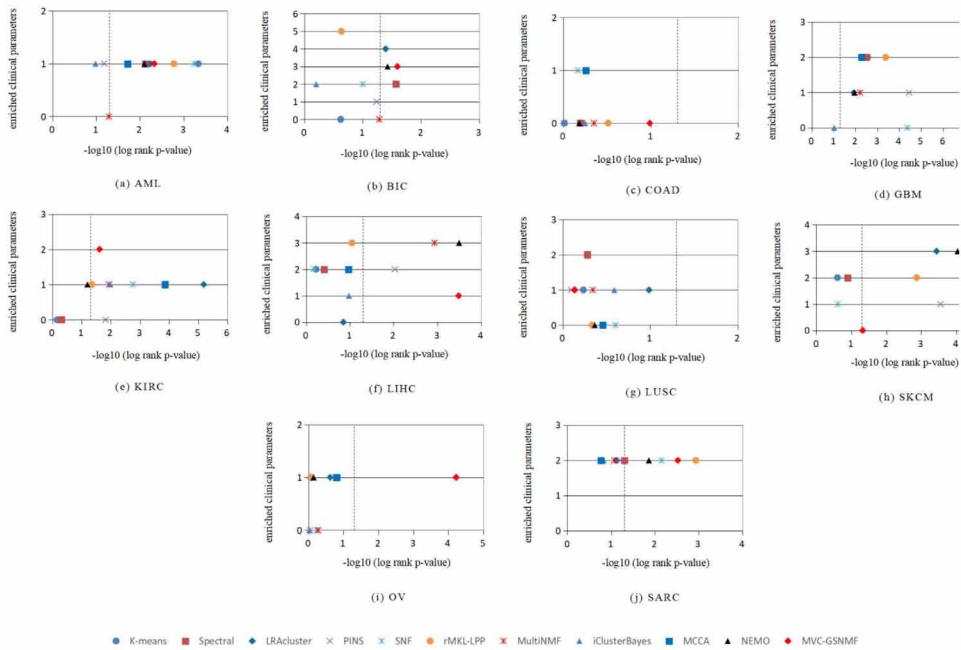


Table 3 presents the number of enriched clinical labels found by the 11 algorithms. Six clinical labels have been selected for significant clinical analysis, such as pathologic stage, age at diagnosis, gender, and pathologic TMN. However, not all cancer data sets have data with all clinical labels. The total number of enriched clinical labels found by each algorithm is shown in Table 3. The rMKL-LPP gained the largest number with 17 enriched labels. NEMO ranked second with 15 enriched labels, while Multi-GSNMF and LRAcluster tied for fourth with 13 enriched labels. However, it is unreasonable to measure the performance of the algorithms by the count of the enriched clinical labels alone, because the number of clinical labels varies greatly among different cancer data sets.

The Kaplan–Meier (KM) survival rate curves of the subtyping identified by Multi-GSNMF on 10 cancer data sets are shown as Figure 5. The subtyping of the proposed algorithm were found to have significantly different survival rates on the 10 cancer data sets except COAD, LUSC, and SKCM, which is the greatest number of compared methods. As we can see in Figure 5(i), Multi-GSNMF suggested that the subtypes of cancer patients in the OV data set be divided into four classes. The KM survival rate curves of the subtyping show significantly different survival rates (p value = $6.0215e^{-5}$).

Table 4 summarizes the performance of the 10 algorithms on the 10 real-world data sets. The numbers of cancer types with significantly different survival rates for each algorithm of 10

Table 3. Number of Enriched Clinical Labels of the 11 Algorithms

Methods	k-means	Spectral	LRA cluster	PINS	SNF	rMKL-LPP	Multi NMF	iCluster Bayes	MCCA	NEMO	Multi-GSNMF
Total	11	14	13	9	10	17	10	11	12	15	13

Figure 5. The KM Survival Rate Curves of the Subtyping by Multi-GSNMF on 10 Cancer Data Sets

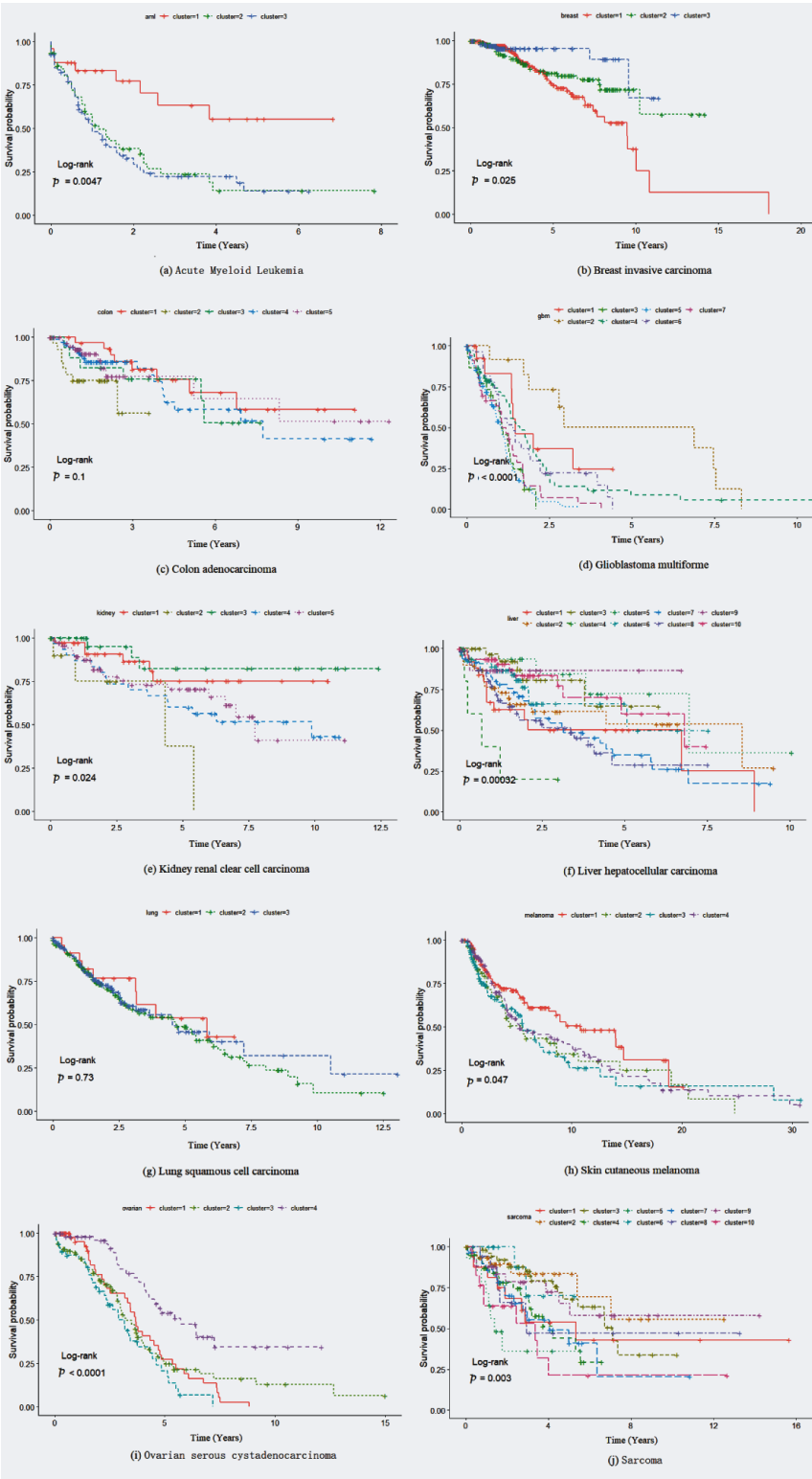


Table 4. Numbers of Data Sets With Significant Results for all the Testing Algorithms

Methods	k-means	Spectral	low-rank approximation based multi-omics data clustering	perturbation clustering for data integration and disease subtyping	similarity network fusion	regularized multiple kernel learning with locality preserving projection	Multi non-negative matrix factorization	iCluster Bayes	Multiset canonical correlation analysis	neighborhood-based multiomics clustering	Multi- graph and sparsity regularized non-negative matrix factorization
Significantly different survival rates	3	4	5	4	4	5	4	2	5	6	7
Enriched clinical labels	7	8	8	7	7	8	6	7	8	8	8

testing data sets are shown in the first row, while the second row is the number of data sets with no less than one enriched clinical label by each algorithm. Multi-GSNMF obtained the highest of the seven cancer types whose clustering results have significantly different prognosis, whereas the 10 other methods found up to six cancer types. Meanwhile, Multi-GSNMF gained eight cancer data sets with no less than one enriched clinical label, which is one of the most among the compared algorithms. In this view, the Multi-GSNMF outperformed the other methods on the 10 testing cancer data sets.

CONCLUSION

A unified multiview clustering algorithm was proposed for cancer subtyping based on MultiNMF. By incorporating the local geometrical structures and sparsity constraints into a unified objective function, Multi-GSNMF showed great advantages in processing bioinformatics data with noise and outliers. Multi-GSNMF is an intermediate integration method based on the NMF dimension reduction. Multi-GSNMF can adaptively obtain the local geometrical structures of each omics data set, which is different from early integration methods without considering the different data distribution in different omics. Compared with late integration methods, an iterative process is used to the continuous optimization of the consensus matrix by Multi-GSNMF, which is beneficial to find the proper low-dimensional representation for each omics data set and consensus matrix for multiomics data sets.

Multi-GSNMF was evaluated on 10 real-world multiomics cancer data sets. While comparing with state-of-the-art multiomics clustering algorithms, Multi-GSNMF obtained cancer subtyping with significantly different survival rates in 7 out of 10 cancer data sets, which was the highest. Multi-GSNMF gained one of the largest numbers of enriched clinical labels on KIRC, while it also achieved the smallest p value of different survival rates on GBM and OV cancer data sets. This result has important guiding implications for cancer subtyping and precise treatment. However, in the bioinformatics data, part of the omics data of some samples are missing. Traditional multiomics clustering methods can only remove samples with the missing parts of omics data. Further research is still needed to develop new approaches to deal with incomplete samples.

FUNDING STATEMENT

This work was supported by the National Natural Science Research Project of China under Grant Nos. 51567002 and 50767001. The authors would like to thank Sujuan Zhou and Faling Yi for their valuable suggestions in the algorithm design. We also thank the reviewers for their efforts to improve the quality of the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

- Amin, J., Sharif, M., Anjum, M. A., Nam, Y., Kadry, S., & Taniar, D. (2021). Diagnosis of COVID-19 infection using three-dimensional semantic segmentation and classification of computed tomography images. *Computers, Materials & Continua*, 68(2), 2451–2467. doi:10.32604/cmc.2021.014199
- Bickel, S., & Scheffer, T. (2004). *Multi-view clustering* [Conference session]. The Fourth IEEE International Conference on Data Mining, Brighton, UK.
- Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press. doi:10.1017/CBO9780511804441
- Cai, D., He, X., Han, J., & Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8), 1548–1560. doi:10.1109/TPAMI.2010.231 PMID:21173440
- Chaudhuri, K., Kakade, S., Livescu, K., & Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. *Proceedings of the 26th International Conference on Machine Learning*.
- Chikhi, N. F. (2016). Multi-view clustering via spectral partitioning and local refinement. *Information Processing & Management*, 52(4), 618–627. doi:10.1016/j.ipm.2015.12.007
- DatasetS. (2009). [Data set]. <http://mlg.ucd.ie/datasets/3sources.html>
- Elhamifar, E., & Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11), 2765–2781. doi:10.1109/TPAMI.2013.57 PMID:24051734
- Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., Van't Veer, L. J., Lopez-Bigas, N., & Stuart, J. M. et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4), 929–944. doi:10.1016/j.cell.2014.06.049 PMID:25109877
- Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5, 1457–1469.
- Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8, 84. doi:10.3389/fgene.2017.00084 PMID:28670325
- Huang, Z., & Wu, J. (2022). A multi-view clustering method with low-rank and sparsity constraints for cancer subtyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6), 3213–3223. doi:10.1109/TCBB.2021.3122917 PMID:34705654
- IARC. (2021). *IARC Biennial Report 2020-2021*. <https://publications.iarc.fr/607>
- Janeja, V. P., Namayanja, J. M., Yesha, Y., Kench, A., & Misal, V. (2020). Discovering similarity across heterogeneous features: A case study of clinico-genomic analysis. *International Journal of Data Warehousing and Mining*, 16(4), 63–83. doi:10.4018/IJDWM.2020100104
- Kumar-Sinha, C., & Chinnaiyan, A. M. (2018). Precision oncology in the age of integrative genomics. *Nature Biotechnology*, 36(1), 46–60. doi:10.1038/nbt.4017 PMID:29319699
- Lê Cao, K., Martin, P. G. P., Robert-Granié, C., & Besse, P. (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics*, 10(1), 34. doi:10.1186/1471-2105-10-34 PMID:19171069
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. doi:10.1038/44565 PMID:10548103
- Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Neural Information Processing Systems (NIPS 2000)*. MIT Press.

- Li, G., Han, K., Pan, Z., Wang, S., & Song, D. (2021a). Multi-view image clustering via representations fusion method with semi-nonnegative matrix factorization. *IEEE Access: Practical Innovations, Open Solutions*, 9, 96233–96243. doi:10.1109/ACCESS.2021.3083501
- Li, R., Wang, X., Quan, W., Song, Y., & Lei, L. (2021b). Robust and structural sparsity auto-encoder with L21-norm minimization. *Neurocomputing*, 425, 71–81. doi:10.1016/j.neucom.2020.02.051
- Li, Y., Wu, F., & Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2), 325–340. doi:10.1093/bib/bbw113 PMID:28011753
- Liang, N., Yang, Z., Li, Z., Sun, W., & Xie, S. (2020). Multi-view clustering by non-negative matrix factorization with co-orthogonal constraints. *Knowledge-Based Systems*, 194, 105582. doi:10.1016/j.knsys.2020.105582
- Liu, J., Wang, C., Gao, J., & Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. *Proceedings of the 2013 SIAM International Conference on Data Mining*.
- Liu, J., Wang, D., Gao, Y., Zheng, C., Xu, Y., & Yu, J. (2018). Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3), 974–987. doi:10.1109/TCBB.2017.2665557 PMID:28186906
- Liu, Y., Gu, Q., Hou, J. P., Han, J., & Ma, J. (2014). A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, 15(1), 37. doi:10.1186/1471-2105-15-37 PMID:24491042
- Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., & Hilsenbeck, S. G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics (Oxford, England)*, 19(1), 71–86. doi:10.1093/biostatistics/kxx017 PMID:28541380
- Nguyen, T., Tagett, R., Diaz, D., & Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, 27(12), 2025–2039. doi:10.1101/gr.215129.116 PMID:29066617
- Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics (London, Ont.)*, 5(2), 111–126. doi:10.1002/env.3170050203
- Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8), 1160–1167. doi:10.1200/JCO.2008.18.1370 PMID:19204204
- Prasad, V., Fojo, T., & Brada, M. (2016). Precision oncology: Origins, optimism, and potential. *The Lancet. Oncology*, 17(2), e81–e86. doi:10.1016/S1470-2045(15)00620-8 PMID:26868357
- Rajinikanth, V., & Kadry, S. (2021). Development of a framework for preserving the disease-evidence-information to support efficient disease diagnosis. *International Journal of Data Warehousing and Mining*, 17(2), 63–84. doi:10.4018/IJDWM.2021040104
- Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Research*, 46(20), 10546–10562. doi:10.1093/nar/gky889 PMID:30295871
- Rappoport, N., & Shamir, R. (2019). NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics (Oxford, England)*, 35(18), 3348–3356. doi:10.1093/bioinformatics/btz058 PMID:30698637
- Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., & Sander, C. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*, 7(4), e35236. doi:10.1371/journal.pone.0035236 PMID:22539962
- Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England)*, 25(22), 2906–2912. doi:10.1093/bioinformatics/btp543 PMID:19759197
- Speicher, N. K., & Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics (Oxford, England)*, 31(12), i268–i275. doi:10.1093/bioinformatics/btv244 PMID:26072491

- Sun, F., Xu, M., Hu, X., & Jiang, X. (2016). Graph regularized and sparse nonnegative matrix factorization with hard constraints for data representation. *Neurocomputing*, 173, 233–244. doi:10.1016/j.neucom.2015.01.103
- Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3), 333–337. doi:10.1038/nmeth.2810 PMID:24464287
- Wang, X., Zhang, T., & Gao, X. (2019). Multi-view clustering based on non-negative matrix factorization and pairwise measurements. *IEEE Transactions on Cybernetics*, 49(9), 3333–3346. doi:10.1109/TCYB.2018.2842052 PMID:29994496
- Wang, Y., Zhou, G., Guan, T., Wang, Y., Xuan, C., Ding, T., & Gao, J. (2022). A network-based matrix factorization framework for ceRNA co-modules recognition of cancer genomic data. *Briefings in Bioinformatics*, 23(5). 10.1093/bib/bbac154
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10), 1113–1120. doi:10.1038/ng.2764 PMID:24071849
- Wen, J., Zhang, B., Xu, Y., Yang, J., & Han, N. (2018). Adaptive weighted nonnegative low-rank representation. *Pattern Recognition*, 81, 326–340. doi:10.1016/j.patcog.2018.04.004
- Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(281), 28. Advance online publication. doi:10.2202/1544-6115.1470 PMID:19572827
- Wu, D., Wang, D., Zhang, M. Q., & Gu, J. (2015). Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification. *BMC Genomics*, 16(1), 1022. Advance online publication. doi:10.1186/s12864-015-2223-8 PMID:26626453
- Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on nonnegative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery. doi:10.1145/860435.860485