Hindawi
Complexity
Volume 2020, Article ID 4865738, 13 pages
https://doi.org/10.1155/2020/4865738

Research Article

Block-Constraint Laplacian-Regularized Low-Rank Representation and Its Application for Cancer Sample Clustering Based on Integrated TCGA Data

Juan Wang[1], **Jin-Xing Liu**[1], **Chun-Hou Zheng**[2], **Cong-Hai Lu**[1], **Ling-Yun Dai**[1] **and Xiang-Zhen Kong**[1]

[1]*School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 276826, China*
[2]*School of Software Engineering, Qufu Normal University, Qufu, Shandong 273165, China*

Correspondence should be addressed to Jin-Xing Liu; sdcavell@126.com

Received 3 May 2019; Revised 13 December 2019; Accepted 2 January 2020; Published 27 January 2020

Academic Editor: Daniela Paolotti

Low-Rank Representation (LRR) is a powerful subspace clustering method because of its successful learning of low-dimensional subspace of data. With the breakthrough of "OMics" technology, many LRR-based methods have been proposed and used to cancer clustering based on gene expression data. Moreover, studies have shown that besides gene expression data, some other genomic data in TCGA also contain important information for cancer research. Therefore, these genomic data can be integrated as a comprehensive feature source for cancer clustering. How to establish an effective clustering model for comprehensive analysis of integrated TCGA data has become a key issue. In this paper, we develop the traditional LRR method and propose a novel method named Block-constraint Laplacian-Regularized Low-Rank Representation (BLLRR) to model multigenome data for cancer sample clustering. The proposed method is dedicated to extracting more abundant subspace structure information from multiple genomic data to improve the accuracy of cancer sample clustering. Considering the heterogeneity of different genome data, we introduce the block-constraint idea into our method. In BLLRR decomposition, we treat each genome data as a data block and impose different constraints on different data blocks. In addition, graph Laplacian is also introduced into our method to better learn the topological structure of data by preserving the local geometric information. The experiments demonstrate that the BLLRR method can effectively analyze integrated TCGA data and extract more subspace structure information from multigenome data. It is a reliable and efficient clustering algorithm for cancer sample clustering.

1. Introduction

Cancer has seriously threatened the health of people all over the world. For cancer patients, timely detection, accurate diagnosis, and effective treatment are vital for saving lives [1]. Cancer classification, as an important prerequisite for early diagnosis and treatment of cancer, has always been a challenging focus in cancer research. Modern medical research shows that the cause of cancer is the variation and mutation in genes, and these gene mutations and abnormalities cause pathological differences in cancer, forming different classifications in clinical diagnosis [2]. Thus, cancer research at the genetic level has received much attention from biologists.

With the advent of postgenome era in bioinformatics research, vast quantities of genomic data are being generated by DNA-microarray and deep-sequencing techniques [3–6]. Because these techniques can concomitantly profile thousands of genes, these genomic expression data produced by these technologies can fully reflect the transcription activity at a certain point, which affords researchers' avenues to understand and study life mechanism in genome-wide range.

The Cancer Genome Atlas (TCGA), as the largest component of International Cancer Genome Consortium

(ICGC), is by far the largest open genome database for cancer. As of the end of the TCGA project, the TCGA database has collected more than 11,000 cancer cases involving 33 cancer types [7]. TCGA project aims to comprehensively and systematically study the biological and molecular basis of the formation, growth, and metastasis of cancer cells by mapping the genome of human cancers. The TCGA database can provide us with diverse genomics data. These genome data provide an unprecedented opportunity for us to systematically and comprehensively consider different genetic aberrations of biological processes. Therefore, cancer research based on TCGA data has become a hotspot in the field of bioinformatics.

Clustering of cancer samples is an important means of cancer classification. Its purpose is to find samples' sample groups with similar expression. Based on TCGA data, a large number of articles on cancer clustering have been produced. For example, Yu et al. developed a method named Graph-based Consensus Clustering (GCC) to research the classes of the samples based on microarray data [8]. Zheng et al. adopted Nonnegative Matrix Factorization (NMF) and sparse NMF methods to study tumor clustering [9]. Based on the maximum correntropy criterion, Wang et al. proposed a new Nonnegative matrix factorization method named NMF maximum correntropy criterion (NMF-MCC) for cancer clustering from gene expression data [10]. Kong et al. presented a P-norm Singular Value Decomposition (PSVD) method for clustering of tumor [11]. Feng et al. enforced graph-Laplacian regularization and P-norm on PCA and presented the PgLPCA method for selecting feature genes and sample clustering [12]. Virmani et al. used DNA methylation data to cluster lung cancer [13]. Ye et al. studied tumor clustering based on independent component analysis (ICA) and affinity propagation (AP) [14]. Based on genomic data, Liu et al. adopted Robust Principal Component Analysis (RPCA) approach to research tumor clustering [15]. Liu et al. presented a network-assisted coclustering method to identify the cancer subtype [16]. These studies show that besides gene expression data, other genomic data in TCGA also contain the feature information needed for cancer clustering and can be used as feature source for cancer clustering research. Therefore, it is reasonable to think that the integrated data composed of multiple genome data can contain more cancer clustering features than the single genome data, which is helpful to study cancer clustering better. However, different genomic data in the TCGA database come from different categories of genomics assays and therefore have different characteristics. In other words, these genomic data are heterogeneous, which makes the integration and analysis of different genome data become a major bottleneck in bioinformatics research [17]. Hence, most cancer clustering methods are based on single genomic data in the TCGA database, more frequently on gene expression data. This may ignore the interaction of different genetic factors, which is not conducive to the detection of cancer pathogenesis [18]. Obviously, these clustering methods cannot be directly used for comprehensive analysis of integrated TCGA data. In this case, how to establish an effective clustering algorithm for comprehensive analysis of TCGA integrated data to further improve the reliability of cancer clustering has become an urgent problem.

In recent years, Liu et al. developed a novel matrix transformation method known as Low-Rank Representation (LRR) method [19] for subspace segmentation. The LRR method is based on an important assumption that the high-dimensional data are approximated as mappings of unknown low-dimensional space. That is, the high-dimensional data can be recovered from the low-dimensional space. Under this assumption, LRR aims at finding the lowest-rank structural representation of each sample through low-rank constraint. And based on the recovered lowest-rank representation matrix, each sample is grouped into its own subspace. In LRR, because the global space information of input data is exploited to recover the subspace structures embedded in the high-dimensional data, LRR can effectively pick up the underlying subspace structures of data. As a result, the LRR method has achieved excellent performance in subspace segmentation and has been frequently applied in many fields [20–26]. It is well known that, in the real world, high-dimensional data often reside on unknown nonlinear manifolds. However, the classical LRR method loses sight of the local structure information in data, resulting in the loss of the inherent topological characteristics of the nonlinear manifold.

Meanwhile, with the deepening of manifold learning theory and graph theory research, more and more researchers introduce the graph regularization constraint into their research algorithms [27–33]. For example, Long et al. presented a graph-regularized discriminative nonnegative matrix factorization (GDNMF) method [29]; in the GDNMF model, the discriminative information and local geometrical information were taken into account by imposing the graph regularization constraint on the NMF model. Huang et al. presented Hypergraph-based Attribute Predictor (HAP) for attribute learning [31]. To further improve the classification performance of Extreme Learning Machine (ELM), Peng et al. proposed a graph-regularized ELM named as GELM [32]. Cheng et al. proposed a Graph-regularized Dual Lasso method to integrate the geometrical structure within traits and genetic markers [33]. Similarly, in order to learn the topological structure of data better, researchers introduced manifold learning into the LRR method [34–38]. For example, in order to improve the effectiveness of facial expression recognition, Wang et al. presented a regularized low-rank representation approach by combining linear subspace learning with data recovery [34]. Yin et al. combined LRR with graph regularizer and developed the Nonnegative Sparse Hyper-Laplacian-regularized LRR (NSHLRR) method [36]. Wang et al. put forward Laplacian-regularized Low-Rank Representation (LLRR) to identify different expression genes [37]. Besides, these LRR-based methods combining graph regularization have also aroused great interest of biologists and been used in bioinformatics modeling for cancer clustering or cancer classification. Gan et al. applied latent low-rank representation to derive features for tumor clustering [39]. Wang et al. proposed Mixed-norm Laplacian regularized Low-Rank Representation (MLLRR) and applied it to tumor clustering [40]. Xia et al.

presented a self-training subspace clustering algorithm under low-rank representation (SSC-LRR) to model gene expression data for cancer classification [41]. Just recently, Wang et al. used the LLRR method to cluster cancer samples based on gene expression data [42]. Although these studies show that these LRR-based methods with manifold constraint have good performance in cancer clustering, the applicability of these methods in multitype integrated data analysis needs further study.

Inspired by the success of the LRR method and graph regularization, in this work, we present a novel method referred to as Block-constraint Laplacian-regularized Low-Rank Representation (BLLRR) to research cancer sample clustering. BLLRR method is devoted to obtaining a lowest-rank representation matrix which reflects the similarity between samples through comprehensive analysis of integrated TCGA data. Considering that different types of TCGA data have different characteristics and noise, in our method, we treat each type of data as a data block and impose different constraint strengths on different types of data. These different parameters can well balance the noise from different genomic data. In additional, in order to maintain the nonlinear geometrical relationships of real data, graph Laplacian based on manifold is introduced into BLLRR. Graph Laplacian, also named graph regularization, can maximize the smoothness of the nonlinear manifold of data by maintaining local geometrical relationships within data, which greatly enhances the capability of the BLLRR method to learn the subspace structure. Our contributions of this paper are listed as follows. (1) A framework of cancer sample clustering based on multigenome data is come up with. This will bring cancer clustering research out of the confinement of analyzing single gene expression data. (2) We develop a novel method called BLLRR to model integrated TCGA data. In the BLLRR method, we introduce the block-constraint idea to decompose integrated TCGA data. Block-constraint solves the bottleneck problem of heterogeneous data integration and analysis by imposing different constraints on different genome data. Besides, in order to smooth the nonlinear manifold structure of data, graph regularization is introduced into BLLRR. Both graph regularization and block-constraint enable our method to pick up the subspace structures embedded in multigenome data well. (3) In BLLRR, adaptive balance parameters are proposed to balance the noise of different types of data. Namely, the constraint strength of each type of data is constantly adjusted with iteration, which greatly reduces the trouble of parameter selection and makes the model more adaptable. (4) BLLRR model is applied to the clustering of cancer samples, and many experiments of cancer clustering are provided. The experimental results substantiate the feasibility of cancer clustering based on integrated multigenome data and also show that the BLLRR method has remarkable reliability and accuracy in cancer sample clustering.

The rest of this paper is organized as follows. In Methodology section, firstly, classical LRR and graph Laplacian are briefly reviewed in 2.1 and 2.2, and then the proposed BLLRR method is elaborate in 2.3. In Section 2.3.1, the objective function of BLLRR is given. In Section 2.3.2, the solving process of the BLLRR method is introduced, and the iteration formulas of the optimal solution are given. In Section 2.3.3, the model of decomposition of multigenome data by BLLRR is established. Also, in Section 2.3.4, the clustering process based on the optimal coefficient matrix obtained by BLLRR is introduced. In Section 3, datasets used for experiments are introduced, and the results and discussions of cancer sample clustering experiments are presented. In Section 4, we conclude the paper.

2. Methodology

2.1. LRR. LRR is a representation-based subspace clustering method. The basic assumption of LRR is to treat high-dimensional data as coming from multiple low-dimensional subspaces, and these subspaces are independent [19]. So, high-dimensional data can be regarded as the mapping of data in these low-dimensional subspaces. Based on this, the LRR method is devoted to calculating the mapping weights of high-dimensional data. The weight matrix is often known as the coefficient matrix or low-rank representation matrix. As the nuclear norm is commonly used to approximate rank operator, the resulting problem of LRR is to solve a convex optimization problem with nuclear norm regularization. Supposing the high-dimensional data matrix is represented by $\mathbf{X}$, of which each column vector represents a data point, the problem of LRR is formulated as

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E}} \quad & \|\mathbf{Z}\|_* + \gamma\|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E}, \end{aligned} \tag{1}$$

where $\mathbf{A}$ is referred to as a dictionary matrix by which the whole low-dimensional space can be linearly spanned, $\mathbf{Z}$ is known as the coefficient matrix corresponding to $\mathbf{A}$, $\|\cdot\|_*$ denotes the nuclear norm, $\|\mathbf{Z}\|_*$ is the summation of the singular values of $\mathbf{Z}$, $\mathbf{E}$ is a noise or perturbation term, $\|\cdot\|_1$ denotes the l_1-norm which is a regularization strategy to produce sparse in matrices, $\|\mathbf{E}\|_1$ is the summation of absolute values of elements in $\mathbf{E}$, and γ is a scalar parameter. After LRR decomposition, the coefficient matrix $\mathbf{Z}$ is obtained from high-dimensional data. Ideally, $\mathbf{A}$ is noiseless, and the coefficient matrix $\mathbf{Z}$ is sparse and symmetric. In general, data matrix $\mathbf{X}$ is selected as the dictionary matrix. So, LRR can be reformulated as

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E}} \quad & \|\mathbf{Z}\|_* + \gamma\|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}. \end{aligned} \tag{2}$$

In such a case, coefficient matrix $\mathbf{Z}$ reflects the mapping relationships between all samples. These mapping relationships are actually the similarities between samples, which can reveal the low-dimensional subspace structure embedded in high-dimensional data. Given $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]$, the column vector $\mathbf{z}_i$ denotes the similarities between the i-th sample and all samples. The more similar the two samples are, the more likely they are to come from a subspace. So, subspace clustering can be implemented based on $\mathbf{Z}$.

2.2. Graph Laplacian. As is known to all, the high-dimensional data observed in the real world usually are located on nonlinear low-dimensional manifolds. Keeping the local geometric structure of data is very important for smoothing the nonlinear manifold structure. Graph Laplacian, as a popular approach to preserve the intrinsic structure embedding in high-dimensional data, is proposed on an essential idea named local invariance proposed by Hadsell et al. [43]. Supposing that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ is the observed data, each column vector of $\mathbf{X}$ is a data sample. These data samples and their neighbors form the local geometric structures of original observed data. In practice, the neighborhood relationship is assumed to be linear [44], i.e., each data sample from a local geometry can be treated as a linear union of its neighbors. So, the linear representation coefficients between data samples can efficiently characterize the local geometric structures. According to this, we construct a k-nearest-neighbor graph $\mathbf{G}$. Here, each data sample is treated as a node, so graph $\mathbf{G}$ is with n nodes. At the same time, we define the weight of each edge connecting two nodes of graph $\mathbf{G}$ as follows:

$$W_{ij} = \begin{cases} 1, & \text{if} \quad \mathbf{x}_i \in N_k\left(\mathbf{x}_j\right) \text{or} \, \mathbf{x}_j \in N_k\left(\mathbf{x}_i\right), \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where W_{ij} is the weight value of edge associating nodes i and j, x_i and x_j are data samples corresponding to node i and j, respectively, and $N_k(\mathbf{x}_i)$ is the set of k-nearest-neighbors of node i. The weights of all edges in graph $\mathbf{G}$ constitute a weight matrix denoted as $\mathbf{W}$. Obviously, the affinity between any two nodes of graph $\mathbf{G}$ can be measured by matrix $\mathbf{W}$. According to the idea of local invariance, the nature assumption in manifold theory is that the affinity relations of data samples in input space should be kept in a new space. That is to say, if data samples are nearby to each other in the intrinsic geometry of observed data, then their mappings on the output low-dimensional manifold are nearby too. The hypothesis can be achieved by neighborhood relationships. In mathematics, the relationship can be formulated as follows:

$$\min_{\{\mathbf{z}\}} \quad \sum_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 W_{ij}, \tag{4}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the representations of $\mathbf{x}_i$ and $\mathbf{x}_j$ under the low-dimensional manifold, respectively. Next, we define a diagonal matrix $\mathbf{S}$ with size $n \times n$, and the i-th diagonal element of $\mathbf{S}$ is defined as $S_{ii} = \sum_j W_{ij}$. Apparently, S_{ii} indicates the total affinities related with sample $\mathbf{x}_i$, so matrix $\mathbf{S}$ is often called the degree matrix. Accordingly, a Laplacian matrix [45] $\mathbf{L}$ is defined as $\mathbf{L} = \mathbf{S} - \mathbf{W}$. It is not difficult to prove that the relationship defined by (4) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{Z}} \sum_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 W_{ij} \\ &\quad = \min_{\mathbf{Z}} tr\left(\mathbf{Z}\left(\mathbf{S} - \mathbf{W}\right)\mathbf{Z}^T\right) \\ &\quad = \min_{\mathbf{Z}} tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^T\right). \end{aligned} \tag{5}$$

Because formulation (5) can describe the local adjacency relation of graph $\mathbf{G}$ by edge weight matrix $\mathbf{W}$ which keeps the affinity between pairs of nodes, it is called graph Laplacian. This rule is essential to preserve the inherent geometric structure of the original data distribution.

2.3. BLLRR Method

2.3.1. Definition of BLLRR. The traditional LRR [19] method and its improved algorithms, such as NSHLRR [36], LLRR [37], and SSC-LRR [41], improve the algorithm robustness to noise by enforcing l_1-norm or $l_{2,1}$-norm constraint on the perturbation item. In these methods, all samples are subject to uniform constraint strength; therefore, these methods are only applicable to the study of a single type of data. For heterogeneous data, these methods cannot be used directly. However, in practice, we need to obtain more useful information through comprehensive analysis of various heterogeneous data. For the analysis of multiple heterogeneous data, there are two issues need to be considered. One is that heterogeneous data have different characteristics because they come from different experiments or environments. The other is multiple heterogeneous data will bring more complex noise. Based on these two aspects, when dealing with multiple heterogeneous data, we introduce the block-constraint idea. Namely, we treat each category of data as a data block, and on different data blocks, we impose different constraint strengths. Block-constraint can not only balance the noise from different data but also preserve the feature information in the data by following the characteristics of heterogeneous data. In addition, similar to LLRR, to well discover the intrinsic geometrical structure embedding in the high-dimensional space, manifold constraint is also introduced into the algorithm. So, the optimization problem is formulated as follows:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E}} \quad & \|\mathbf{Z}\|_* + \alpha\|\mathbf{Z}\|_1 + \lambda tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathbf{T}}\right) + \sum_{l=1}^{c} \gamma_l \left\|\mathbf{E}_l\right\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \; \mathbf{Z} > 0, \end{aligned} \tag{6}$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_c) \in R^{m \times n}$ is the input data matrix that is a collection of multiclass data, where c is the number of data categories and $\mathbf{X}_l$ is the l-th category data. Accordingly, $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_c) \in R^{m \times n}$ is the noise matrix, to be specific, $\mathbf{E}_l$ is the noise signal with regard to $\mathbf{X}_l$. α and λ are penalty parameters. $\gamma_l\,(l = 1, \ldots, c)$ is the weighting parameter to balance the noise item of different categories. In (6), the LRR method is combined with graph Laplacian and block-constraint, so it is named as the Block-constraint Laplacian regularized Low-Rank Representation method. Obviously, when $c = 1$, the BLLRR model degenerates into the LLRR model whose objective function is as follows:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E}} \quad & \|\mathbf{Z}\|_* + \alpha\|\mathbf{Z}\|_1 + \lambda tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathbf{T}}\right) + \gamma\|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \; \mathbf{Z} > 0. \end{aligned} \tag{7}$$

2.3.2. The Optimization of BLLRR. In order to recover the low-rank representation from data, many algorithms have been developed [46–48]. Specially, the ADM with Linearized Adaptive Penalty (LADMAP) [48] is a more efficient algorithm. In this paper, LADMAP is also applied to resolve problem (6).

Firstly, an auxiliary variable $\mathbf{J}$ is introduced to make problem (6) separable. So, equation (6) can be converted to the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E},\mathbf{J}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{J}\|_1 + \lambda tr\left(\mathbf{ZLZ}^{\mathbf{T}}\right) + \sum_{l=1}^{c}\gamma_l\left\|\mathbf{E}_l\right\|_1\\ &\text{s.t.}\mathbf{X} = \mathbf{XZ} + \mathbf{E},\ \mathbf{Z} = \mathbf{J},\quad \mathbf{Z} > 0.\end{aligned} \tag{8}$$

Then, we remove the linear constraints in (8) by introducing the augmented Lagrangian formulation. Therefore, optimization problem (8) can be transformed into the following:

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E},\mathbf{J}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{J}\|_1 + \lambda tr\left(\mathbf{ZLZ}^{\mathbf{T}}\right) + \sum_{l=1}^{c}\gamma_l\left\|\mathbf{E}_l\right\|_1\\ &\quad + \langle\mathbf{M}_1, \mathbf{X} - \mathbf{XZ} - \mathbf{E}\rangle + \langle\mathbf{M}_2, \mathbf{Z} - \mathbf{J}\rangle\\ &\quad + \frac{\mu}{2}\left(\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2\right),\end{aligned} \tag{9}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are Lagrangian multipliers, μ is a penalty parameter that can be adaptively adjusted, $\|\cdot\|_F$ is the matrix Frobenius norm, and the value of $\|\mathbf{Y}\|_F$ is the sum of squares of all elements in matrix $\mathbf{Y}$.

Finally, in order to optimize the variables $\mathbf{Z}$, $\mathbf{J}$, and $\mathbf{E}$ by alternate updating, the original optimization problem is divided into three subproblems:

$$\begin{aligned}\Lambda_1 &= \|\mathbf{Z}\|_* + \lambda tr\left(\mathbf{ZLZ}^{\mathbf{T}}\right) + \langle\mathbf{M}_1, \mathbf{X} - \mathbf{XZ} - \mathbf{E}\rangle\\ &\quad + \langle\mathbf{M}_2, \mathbf{Z} - \mathbf{J}\rangle + \frac{\mu}{2}\left(\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2\right)\\ &= \|\mathbf{Z}\|_* + \lambda tr\left(\mathbf{ZLZ}^{\mathbf{T}}\right) + \frac{\mu}{2}\left\|\mathbf{X} - \mathbf{XZ} - \mathbf{E} + \frac{1}{\mu}\mathbf{M}_1\right\|_F^2\\ &\quad + \frac{\mu}{2}\left\|\mathbf{Z} - \mathbf{J} + \frac{1}{\mu}\mathbf{M}_2\right\|_F^2,\end{aligned} \tag{10}$$

$$\begin{aligned}\Lambda_2 &= \alpha\|\mathbf{J}\|_1 + \langle\mathbf{M}_2, \mathbf{Z} - \mathbf{J}\rangle + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{J}\|_F^2 = \alpha\|\mathbf{J}\|_1\\ &\quad + \frac{\mu}{2}\left\|\mathbf{Z} - \mathbf{J} + \frac{1}{\mu}\mathbf{M}_2\right\|_F^2,\end{aligned} \tag{11}$$

$$\begin{aligned}\Lambda_3 &= \sum_{l=1}^{c}\gamma_l\left\|\mathbf{E}_l\right\|_1 + \langle\mathbf{M}_1, \mathbf{X} - \mathbf{XZ} - \mathbf{E}\rangle + \frac{\mu}{2}\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2\\ &= \sum_{l=1}^{c}\gamma_l\left\|\mathbf{E}_l\right\|_1 + \frac{\mu}{2}\left\|\mathbf{X} - \mathbf{XZ} - \mathbf{E} + \frac{1}{\mu}\mathbf{M}_1\right\|_F^2.\end{aligned} \tag{12}$$

(1) The Computation of $\mathbf{Z}$. Fixed $\mathbf{E}$ and $\mathbf{J}$, the iteration formula of $\mathbf{Z}$ can be obtained by solving subproblem (10).

Firstly, we define a quadratic term as follows:

$$\begin{aligned}Q(\mathbf{Z}, \mathbf{E}, \mathbf{J}, \boldsymbol{\mu}, \mathbf{M}_1, \mathbf{M}_2) &= \lambda tr\left(\mathbf{ZLZ}^{\mathbf{T}}\right) + \frac{\mu}{2}\left\|\mathbf{X} - \mathbf{XZ} - \mathbf{E} + \frac{1}{\mu}\mathbf{M}_1\right\|_F^2\\ &\quad + \frac{\mu}{2}\left\|\mathbf{Z} - \mathbf{J} + \frac{1}{\mu}\mathbf{M}_2\right\|_F^2.\end{aligned} \tag{13}$$

Then, subproblem (10) is recast as the following objective function:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \langle\nabla_{\mathbf{Z}}Q(\mathbf{Z}_K), \mathbf{Z} - \mathbf{Z}_K\rangle + \frac{\mu_K\eta}{2}\|\mathbf{Z} - \mathbf{Z}_K\|_F^2, \tag{14}$$

where $\nabla_z Q(\mathbf{Z}_K) = \lambda(\mathbf{Z}_K\mathbf{L}^T + \mathbf{Z}_K\mathbf{L}) + \mu_K(\mathbf{Z}_K - \mathbf{J}_K + \mathbf{M}_2^K/\mu_K) + \mu_K\mathbf{X}^T(\mathbf{XZ}_K - \mathbf{X} + \mathbf{E}_K - \mathbf{M}_1^K/\mu_K)$, $\eta = 2\lambda\|\mathbf{L}\|_2 + \mu_K(1 + \|\mathbf{X}\|_2^2)$.

Finally, the solution of $\mathbf{Z}$ is given by

$$\mathbf{Z}_{K+1} = \Theta_{1/\eta\mu_K}\left(\mathbf{Z}_K - \frac{\nabla_{\mathbf{Z}}Q(\mathbf{Z}_K)}{\eta}\right), \tag{15}$$

where $\Theta(\cdot)$ is an operator of singular value threshold [49] and $\Theta_\varepsilon(A)$ is defined as $\Theta_\varepsilon(A) = US_\varepsilon(\sum)V^T$, in which ε is a threshold, $S(\cdot)$ is a shrinkage operator, and $S_\varepsilon(x)$ is defined as $S_\varepsilon(x) = \text{sgn}(x)\max(|x| - \varepsilon, 0)$, where $\text{sgn}(\cdot)$ is a symbolic operator.

(2) The Computation of $\mathbf{J}$. Fixed the current value of other variables, the iteration formula of $\mathbf{J}$ can be obtained by solving subproblem (11). The solution of $\mathbf{J}$ is given by

$$\mathbf{J}_{K+1} = \max\left\{\Omega_{\alpha/\mu_K}\left(\mathbf{Z}_{K+1} + \frac{\mathbf{M}_2^K}{\mu_K}\right), 0\right\}, \tag{16}$$

where $\Omega(\cdot)$ is an operator of soft shrinkage and $\Omega_\varepsilon(x)$ is defined as $\Omega_\varepsilon(x) = \max(x - \varepsilon, 0) + \min(x + \varepsilon, 0)$.

(3) The Computation of $\mathbf{E}$. Similarly, fixed $\mathbf{Z}$ and $\mathbf{J}$, the iteration formula of $\mathbf{E}$ can be obtained by solving subproblem (12). According to Lemma 1 [50], an operator solving subproblem (12) is denoted as $\Gamma(\cdot)$. So, the solution of $\mathbf{E}$ is as follows:

$$\Gamma_\varsigma(\mathbf{E}_l) = \mathbf{D}_l\max\left(0, 1 - \frac{\varsigma}{\|\mathbf{D}_l\|_F}\right) = \begin{cases}\dfrac{\|\mathbf{D}_l\|_F - \varsigma}{\|\mathbf{D}_l\|_F}\mathbf{D}_l, & \text{if}\|\mathbf{D}_l\|_F > \varsigma,\\ 0, & \text{otherwise,}\end{cases} \tag{17}$$

for $l = 1, \ldots, c$.

Here, $\mathbf{E}_l$ is the l-th submatrix of $\mathbf{E}$ and denotes the noise signal corresponding to $\mathbf{X}_l$. $\mathbf{D} = \mathbf{X} - \mathbf{XZ}_{K+1} + \mathbf{M}_1^K/\mu_K$, and $\mathbf{D}_l$ is the l-th submatrix of $\mathbf{D}$. $\varsigma = \gamma_l/\mu_K$ denotes the threshold of the corresponding block.

The iteration formulas of $\mathbf{M}_1$, $\mathbf{M}_2$, and μ are as follows:

$$\mathbf{M}_1^{\mathrm{K}+1} = \mathbf{M}_1^{\mathrm{K}} + \mu_{\mathrm{K}}(\mathbf{X} - \mathbf{XZ}_{\mathrm{K}+1} + \mathbf{E}_{\mathrm{K}+1}), \tag{18}$$

$$\mathbf{M}_2^{\mathrm{K}+1} = \mathbf{M}_2^{\mathrm{K}} + \mu_{\mathrm{K}}(\mathbf{Z}_{\mathrm{K}+1} - \mathbf{J}_{\mathrm{K}+1}), \tag{19}$$

$$\mu_{K+1} = \min\left(\mu_{\max}, \rho_K \mu_K\right), \tag{20}$$

where $\rho_K = \left\{ \begin{matrix} \rho_0, & \text{if } \mu_K \cdot \max\{\eta\|\mathbf{Z}_K - \mathbf{Z}_{K+1}\|, \|\mathbf{J}_K - \mathbf{J}_{K+1}\|, \|\mathbf{E}_K - \mathbf{E}_{K+1}\|\} \le \varepsilon \\ 1, & \text{otherwise} \end{matrix} \right\}$.

The main procedure of BLLRR is shown in Algorithm 1.

2.3.3. The BLLRR Model of Integrated TCGA Data. Though people have been studying cancer clustering based on the gene expression for many years, it has been increasingly recognized that DNA copy number variation and DNA methylation also play important role in cancer understanding and clustering research [51–54]. Moreover, as mentioned earlier, TCGA dataset can provide a variety of genomic data for each sample, which make it possible to study cancer based on a variety of biological processes. Therefore, we integrate these different genomics data as an integrated feature source to research cancer clustering. Figure 1 shows a schematic diagram of the multiassay genomic data. In Figure 1, mRNA expression, DNA copy number, and DNA methylation represent different genomics assay data from TCGA, in which each row represents a feature from a certain type of genome data, and each column represents a sample. Therefore, in the integrated data, each sample contains all the features from three categories of genomic data.

Now, we focus on integrated multigenome data. In our integrated data, there are three different types of genome data. And each category data is regarded as a data block. Because of the heterogeneity of different data blocks, in the BLLRR method, we impose different constraints on each data block, which are called as block-constraint. After BLLRR decomposition, the coefficient matrix $\mathbf{Z}$, which reflects the similarity between samples, is obtained. It is not difficult to understand that the samples with high similarity can be regarded as located in the same subspace. Consequently, based on $\mathbf{Z}$, the samples can be clustered. The schematic depiction of BLLRR decomposition of integrated multigenome data is shown as Figure 2. In this figure, $\mathbf{X}$ is the multigenome data matrix, $\mathbf{Z}$ is the low-rank representation matrix, $\mathbf{E}$ is the noise matrix, and γ_l is the constraint intensity on the l-th category data.

As shown in Figure 2, the observation data are decomposed into two parts: one is the low-rank matrix and the other is the noise matrix. Of course, an appropriate restraint strength, i.e., scale parameter γ, is critical to enhance the robustness of BLLRR and obtain accurate similarity patterns between samples. Due to the different constraints imposed on different types of data blocks, it is difficult to tune parameter γ by following the traditional method of parameter tuning. Furthermore, because different types of data have different noises, it is reasonable to think that the noise of a certain type of data is only related to this kind of data. Thus, we propose a new idea called parameter self-regulation to set these parameters γ_l for different data blocks. Specifically, the parameters are adjusted with the iteration process. For the category l, the parameter γ_l is set as follows:

$$\gamma_l^i = \frac{\left\|\mathbf{D}_l^i\right\|_F}{\left\|\mathbf{D}_l\right\|_F}, \tag{21}$$

where γ_l^i is the constraint intensity of the$i$-th feature in the category l. As previously described, $\mathbf{D} = \mathbf{X} - \mathbf{X}\mathbf{Z}_{K+1} + \mathbf{M}_1^K/\mu_K$ is an intermediate matrix generated in the iteration process, and it has the same data dimension and corresponding data block relationship with $\mathbf{E}$. So, $\mathbf{D}_l$ is the matrix corresponding to the category l, and $\mathbf{D}_l^i$ denotes the i-th feature vector. As can be seen from formula (21), in the BLLRR method, we impose different constraints on each feature vector to balance the noise item of different categories of data. And the constraint intensity of each feature vector is calculated by the ratio of the F-norm of feature vector to the F-norm of the data block matrix in which the feature is located. In the iteration process of the BLLRR algorithm, $\mathbf{D}$ is constantly updated, so the constraint strength of each type of data is also constantly adjusted with iteration.

2.3.4. Clustering with BLLRR. As discussed previously, the coefficient matrix $\mathbf{Z}$ obtained after BLLRR decomposition reflects the similarities between samples. According to $\mathbf{Z}$, the samples with high similarity are clustered into one class. However, the observation data from real world are inevitably noisy, so $\mathbf{Z}$ is usually neither sparse nor symmetric. Before using $\mathbf{Z}$ to implement clustering, we need to do some processing on $\mathbf{Z}$ to improve the accuracy of clustering and increase the interpretability of clustering. Firstly, $\mathbf{Z}$ is normalized by rows and shrinked under the appropriate threshold ζ that is very small and close to zero. After the above treatment, $\mathbf{Z}$ becomes a sparse matrix $\widetilde{\mathbf{Z}}$. That is, each sample is similar to only a few other samples, which is critical for clustering problem. Next, we construct an affinity graph using all the samples. Based on $\widetilde{\mathbf{Z}}$, we define an affinity matrix $\widehat{\mathbf{Z}}$ to denote the affinities between samples in the affinity graphs. In $\widehat{\mathbf{Z}}$, both element $\widehat{z}_{ji}$ and $\widehat{z}_{ij}$ denote the affinity of sample i and j, so $\widetilde{z}_{ji}$ is equal to $\widetilde{z}_{ij}$, and $\widetilde{\mathbf{Z}}$ is a symmetric matrix. Consequently, the affinity matrix $\widehat{\mathbf{Z}}$ is defined as $\widehat{\mathbf{Z}} = (|\widetilde{\mathbf{Z}}| + |(\widetilde{\mathbf{Z}})^T|)/2$. So far, based on the affinity matrix, the sample clustering problem can be regarded as a graph segment problem. After the above two steps of processing, the affinity matrix becomes sparse and symmetrical. However, the affinity matrix does not have the block structure needed for clustering and cannot directly obtain the clustering results of samples. Finally, a classical spectral clustering method—K-means is adopted to obtain the final clustering label of the samples based on $\widehat{\mathbf{Z}}$.

The main clustering procedure of BLLRR is shown in Algorithm 2.

3. Experimental Results and Discussion

Firstly, the original datasets from TCGA and their integrated datasets for experiments are introduced. Then, based on experimental datasets, we carry out cancer sample clustering experiments to test the effectiveness of our method. In

Input: Observation matrix $\mathbf{X}$, Laplacian matrix $\mathbf{L}$
Parameter α, λ
Output: $\mathbf{Z}$
Initial: $\mathbf{Z}_0 = \mathbf{E}_0 = \mathbf{J}_0 = \mathbf{M}_1^0 = \mathbf{M}_2^0 = 0, \mu_0 = 10^{-3}, \rho_0 = 2.5, \mu_0 = 10^{-3}, \mu_{\max} = 10^6, \varepsilon = 10^{-2}, \eta = 1.25 \times \|\mathbf{X}\|^2,$
Loop until convergence
  Updating $\mathbf{Z}_{K+1}$ as (15)
  Updating $\mathbf{J}_{K+1}$ as (16)
  Updating $\mathbf{E}_{K+1}$ as (17)
  Updating $\mathbf{M}_1^{K+1}$ as (18)
  Updating $\mathbf{M}_2^{K+1}$ as (19)
  Updating μ_{K+1} as (20)
End Loop

Algorithm 1: LADMAP for solving (9).

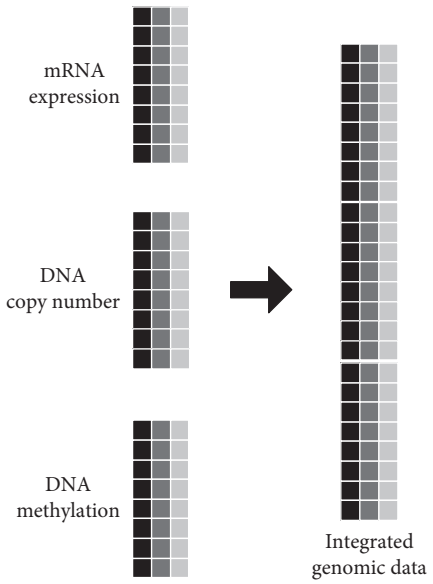

Figure 1: The schematic diagram of integrated multigenome data.

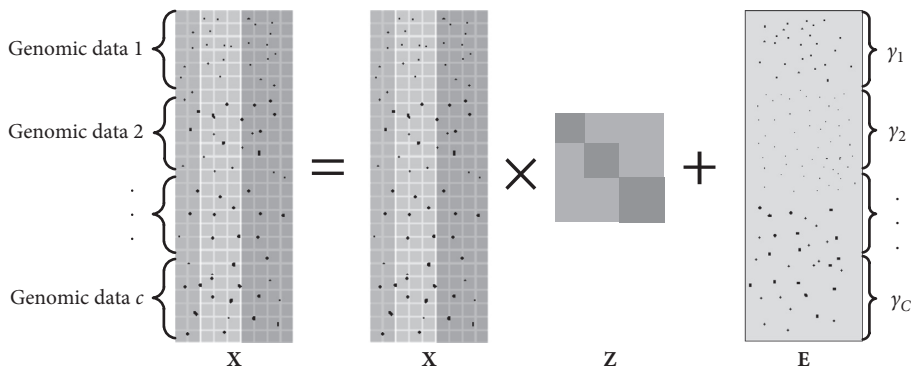

Figure 2: The schematic diagram of BLLRR decomposition of integrated multigenome data.

addition, in order to further demonstrate the performance of BLLRR, we choose K-means, GNMF [27], gLPCA [55], LRR [19], and LLRR [37] as comparison methods in our experiments. In the following section, we give experimental results and discuss the clustering performance of the BLLRR method in detail.

Input: Observation data $\mathbf{X}$, clustering number K
Output: $\widehat{\mathbf{Z}}$
(1) Get the coefficient matrix $\mathbf{Z}$ of problem (8) using BLLRR method.
(2) Normalize $\mathbf{Z}$ by rows as $\mathbf{z}_i = \mathbf{z}_i / \|\mathbf{z}_i\|_2$.
(3) Shrink $\mathbf{Z}$ to get the sparse matrix $\widetilde{\mathbf{Z}}$ by $\widetilde{z}_{ij} = \begin{cases} z_{ij} & \text{if } z_{ij} \geq \zeta \\ 0 & \text{otherwise} \end{cases}$.
(4) Compute the symmetrical affinity matrix by $\widehat{\mathbf{Z}} = (|\widetilde{\mathbf{Z}}| + |(\widetilde{\mathbf{Z}})^T|)/2$.
(5) Adopt K-means to get the cluster label of each sample based on $\widehat{\mathbf{Z}}$.

Algorithm 2: Clustering with BLLRR.

3.1. Dataset. The genomic data used in our experiment are from TCGA. Here, we download three publicly published cancer datasets: Colon Adenocarcinoma (COAD) dataset, Esophagus Cancer (ESCA) dataset, and Head and Neck cancer (HNSC) dataset. Each dataset contains two types of sample labels. One is normal, and the other is tumor. In the COAD dataset, there are 262 tumor samples and 19 normal samples. In the ESCA dataset, there are 183 tumor samples and 9 normal samples. In the HNSC dataset, there are 398 tumor samples and 20 normal samples. So, the total number of samples in the three datasets is 281, 192, and 418, respectively. In addition, each dataset includes three categories of genome data: DNA copy number variation, mRNA expression level, and DNA methylation. Also, in the three datasets, each sample from the same category of genome data contains the same number of genes. Specifically, in DNA copy number data, one sample contains 23,627 genes. In mRNA expression data, one sample contains 20,502 genes. And in DNA methylation data, one sample contains 21,031 genes.

As stated earlier, besides mRNA expression data, both DNA copy number data and DNA methylation data also play important role in cancer clustering research. According to Figure 1, we integrate the three types of genome data from each dataset into multigenome data for cancer sample clustering. The three integrated data are COInteg corresponding to the COAD dataset, ESInteg corresponding to the ESCA dataset, and HNInteg corresponding to the HNSC dataset. Thus, COInteg contains 281 samples and each sample contains 65,160 genes, ESInteg contains 192 samples and each sample contains 65,160 genes, and HNInteg contains 418 samples and each sample contains 65,160 genes.

3.2. Evaluation Index of Clustering Performance. In clustering research, evaluation is a necessary work. Many indexes have been designed to evaluate the performance of the clustering algorithm, such as accuracy (AC), true positive rate (TPR), false positive rate (FPR), receiver operating characteristic (ROC) curve, precision, and F1-measure. In this paper, we use AC, TPR, and FPR to evaluate the clustering performance of the BLLRR algorithm. Next, we will introduce them concisely.

3.2.1. AC. For a given dataset, the ratio of the number of samples correctly clustered to the total number of samples is defined as AC [56]. In practice, AC is calculated by comparing the clustering labels and real labels of samples. The mathematical definition of AC is as follows:

$$\mathrm{AC} = \frac{\sum_{i=1}^{N} \delta(s_i, \mathrm{map}(r_i))}{N}, \tag{22}$$

where N is the total number of samples contained in each experimental dataset, r_i is the clustering label of sample i assigned by the clustering algorithm, s_i is the real label of sample i, and $\delta(s_i, \mathrm{map}(r_i))$ is a function that compares the clustering label of a sample with its real label and gets the result of the comparison. If the clustering label is consistent with the real label, the function value is 1; otherwise, the value is 0. And $\mathrm{map}(r_i)$ is a mapping function that matches the clustering label of the sample to its real label to facilitate label comparison. By the Kuhn–Munkres method [57], the best matching can be achieved.

3.2.2. TPR and FPR. TPR and FPR, as common metrics widely used to evaluate clustering quality, are all calculated based on the confusion matrix. So, let us start with a brief introduction to confusion matrices. Confusion matrix, also known as the error matrix, is a standard format for evaluating. Obfuscation matrix is a two-dimensional matrix. Each row represents an actual class, and each column represents a predicted class. The confusion matrix of a simple case with two classes is shown in Table 1. Generally, among these two classes, the one we are concerned with is designated as a positive class and the other as a negative class. In this table, true positive (TP) denotes the number of positive class samples that are correctly clustered into positive class. True negative (TN) indicates the number of negative class samples that are correctly clustered into negative class. False positive (FP) denotes the number of negative class samples that are incorrectly clustered into positive class. False negative (FN) means the number of positive class samples that are incorrectly clustered into negative class. TPR and FPR are defined as follows:

$$\begin{aligned} \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \end{aligned} \tag{23}$$

From the calculation formulas of TPR and FPR, we can see that the TPR represents the ratio of the number of

Table 1: The confusion matrix with two clusters.

		Assigned class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

samples correctly clustered into the positive class to the total number of samples in the positive class, and the FPR represents the ratio of the number of samples incorrectly clustered into the positive class to the total number of samples in the negative class.

3.2.3. Experimental Results. In this section, based on experimental datasets, many sample clustering experiments are performed to fully demonstrate the performance of our method.

Firstly, we apply LLRR to cluster cancer samples based on DNA copy number variation, mRNA expression level, DNA methylation, and their integrated data. As mentioned earlier, when the BLLRR method is applied to single genomic data, the BLLRR model is equivalent to the LLRR model. The accuracies of the cluster are shown in Table 2. In Table 2, DNA copy number variation is denoted by CN, mRNA expression level is denoted by GE, and DNA methylation is denoted by ME. And the best result on each dataset is shown in bold.

From Table 2, we can see that the clustering accuracy of each single genome data from our three experimental datasets is over 92%. This indicates that each of the three categories of genomic data contains useful information for cancer sample clustering. Next, we compare the clustering results on different genomic data of each dataset. Table 2 shows that, for the COAD dataset and the ESCA dataset, the clustering accuracy on GE data is the best, reaching 95.35% and 96.51%, respectively. And for the HNSC dataset, the clustering accuracy on ME data is the best, reaching 97.22%. This comparison further indicates that, besides GE data, CN data and ME data can also be used as feature source data to study the clustering of cancer samples. At last, for each dataset, we compare the clustering accuracy on integrated multigenome data with that on the single genome data. It is not difficult to see that, on all three datasets, the clustering effect of integrated data is worse than the best clustering effect achieved on single genome data. The fundamental reason for this result is that the LLRR method ignores the heterogeneity of different genome data and imposes the same constraint intensity on integrated multigenome data. So, when LLRR is used to decompose multigenome data, the noise and characteristic information of different genome data cannot be well processed. Obviously, the LLRR model is only suitable for single genome data but not for multigenome data. Summing up the above analysis, we come to the following two conclusions: (1) DNA copy number variation, mRNA expression level data, and DNA methylation are of great significance to the clustering of cancer samples, so it is reasonable to integrate them into multigenome data for cancer sample clustering. (2) When processing integrated multigenome data, the heterogeneity of data must be fully considered.

Secondly, in order to test the clustering performance of the BLLRR method based on multigenome data, the cancer sample clustering experiments are conducted on the three integrated multigenome data. As comparison methods, K-means, GNMF, gLPCA, LRR, and LLRR are also used to cluster cancer samples. Moreover, for the sake of the comparability of the experimental results, we uniformly use K-means algorithm to get the final clustering results for GNMF, gLPCA, LRR, and LLRR, just like the BLLRR method. As we all know, because K-means will randomly select cluster centers for each clustering, when clustering with K-means, there is a small difference in each clustering result. In order to reduce the impact of this difference on the evaluation of experimental results, in all our experiments, we take the average of 30 clustering results as the final result. To be specific, for GNMF, gLPCA, LRR, LLRR, and BLLRR, firstly, we decompose the experimental data and get a matrix for clustering. Then, we use K-means to repeat clustering 30 times based on the obtained matrix and take the mean of 30 times clustering accuracies as the final clustering result. Table 3 gives the clustering accuracy of each method on multigenome data in detail. Similarly, for each dataset, the best result is displayed in bold.

Of these methods used for comparison, LRR and LLRR are LRR-based clustering methods; K-means, GNMF, and gLPCA are traditional methods. Firstly, as can be seen from Table 3, the clustering accuracies of LRR and LLRR are higher than those of three traditional methods on the whole. This benefits from the successful learning of the subspace structure embedded in data by LRR-based methods, which reflects the importance of the subspace structure for clustering research. Secondly, comparing LRR with LLRR, we can see that the clustering performance of LLRR is better than that of LRR. This is due to the introduction of the graph regularization term in the LLRR method. As introduced previously, graph regularization can preserve the geometrical relationships of data and furthermore smooth the nonlinear manifold. Therefore, LLRR has better ability to learn the subspace structure than LRR. Thirdly, we compare BLLRR with LLRR. It is very clear that, on each integrated data, the clustering accuracy of BLLRR is higher than that of LLRR. For LLRR and BLLRR, their basic clustering ideas are consistent. Furthermore, in both algorithms, graph Laplacian is introduced to help better obtain the subspace structure of data. The main difference between the two methods is that when decomposing multigenome data, the idea of block-constraint is introduced into the BLLRR method. In the BLLRR method, each category of genome data contained in the integrated data is regarded as a data block, and different constraints are imposed on different data blocks. Because block-constraint considers the peculiarities of different genome data in multigenome data, it can improve the robustness of BLLRR to complex noise from multigenome data and protect the feature information of each genome data well. However, in the LLRR method, the integrated multigenome data are regarded as single genome data and imposed on a uniform constraint strength, which

Table 2: The clustering accuracy of LLRR on single genome data and integrated multigenome data.

Dataset	COAD				ESCA				HNSC			
	CN	GE	ME	COInteg	CN	GE	ME	ESInteg	CN	GE	ME	HNInteg
LLRR	92.88	**95.35**	94.19	95.04	94.79	**96.51**	94.79	96.39	94.98	96.28	**97.22**	96.31

Table 3: The clustering accuracy on multigenome datasets.

Multigenome data	K-means	GNMF	gLPCA	LRR	LLRR	BLLRR
COInteg	86.99	81.85	93.70	93.59	95.04	**98.56**
ESInteg	96.35	96.35	94.80	95.83	96.39	**96.88**
HNInteg	82.34	84.99	86.82	94.98	96.31	**97.58**

ignores the peculiarities of different types of data. So, BLLRR can deal with multiple heterogeneous data more effectively than LLRR. Finally, comparing the results of BLLRR shown in Table 3 with the results of LLRR based on single genomic data shown in Table 2, we can see that, on all the three datasets, the clustering results of BLLRR on multigenome data are better than the best results of LLRR on single genomic data. This indicates that multigenome data contain more subspace structure information than single genome data and can be used as comprehensive feature source for cancer research. Meanwhile, it again illustrates that BLLRR is capable of mining more useful subspace information from multiple genomic data for sample clustering. Based on the above analysis, we can conclude that the BLLRR method has powerful ability to learn the intrinsic subspace structure within multiple heterogeneous data and can effectively cluster cancer samples by decomposing multiple genomic data.

Now, we would like to further explain the importance of parameter γ_l and the rationality of our setting of γ_l. Firstly, as can be seen from formula (21), we set the corresponding parameter γ_l according to the overall expression level of different genomic data, which helps to set up appropriate constraints for each genome data with different expression levels. Moreover, γ_l will be continuously updated in the iteration. So, parameter γ_l will help to process the complex noises in multigenomic data better. Then, we compare the experimental results of the BLLRR method and the LLRR method to illustrate the rationality of parameter γ_l. As discussed earlier, when a uniform constraint strength is applied to multiple genome data, BLLRR degenerates into LLRR. From the comparative analysis of Tables 2 and 3, we can get the following two points. One is that, for the LLRR method, the clustering results on multigenomic data are worse than those on single genome data. This indicates that it is not feasible to impose uniform constraints on multigenomic data to deal with different noise levels. Second, for the BLLRR method, its clustering result on multigenomic data is better than that on single genome data. This proves that parameter γ_l can effectively balance the complex noises in different genomic data. Summarizing the above analysis, both the formula and the experimental results show that the parameter γ_l obtained in BLLRR is reasonable and effective.

However, the samples in our experimental datasets are extremely imbalanced, that is, there are more tumor samples and fewer normal samples. Sample imbalance is a common problem in the field of bioinformatics. In order to indicate the degree of sample imbalance, for each integrated data, we calculate the ratios of two types of samples, as shown in Table 4. In Table 4, n_T and n_N represent the number of tumor samples and the number of normal samples, respectively. So, n_T/n_N denotes the ratio of tumor samples to normal samples, and n_N/n_T denotes the ratio of normal samples to tumor samples. In this case, the normal samples are surrounded by a large number of tumor samples, which is disadvantageous to the clustering of normal samples.

Finally, in view of this situation, we use TPR and FPR as evaluation measures to research the clustering effect of each class of samples. In cancer clustering research, researchers tend to pay more attention to disease samples, that is, cancer samples or tumor samples. Therefore, we regarded cancer samples as positive samples and normal samples as negative samples. The values of TPR and FPR on all multigenome data are recorded in Table 5. According to the definition of TPR, the larger the value of TPR, the better the clustering effect of cancer samples. And for FPR, the smaller the value of FPR is, the better the clustering effect of normal samples is. So, in Table 5, for each data, both the maximum values of TPR and the minimum values of FPR are remarked in bold. And for ease of comparison, we also use histograms to illustrate the results as shown in Figures 3 and 4.

In our data, because positive class samples are far more than negative class samples, in the following description, positive class samples are also called majority class samples and negative class samples are also called minority class samples. From Figure 3, we can find that the PTR values of various methods are generally high on all three data, especially on ESInteg, the mean value of PTR exceeds 99%. In addition, as can be seen from Figure 4, most FPR values exceed 60%. Especially, from Table 5, we also see that the FPR values of GNMF on COInteg and LRR on HNInteg are 100%. These results show that the extreme imbalance of sample distribution is beneficial to the clustering of majority class samples, but it is a great challenge to the clustering of minority class samples. In order to demonstrate the clustering performance of BLLRR for minority class samples, we compare LRR, LLRR, and BLLRR. Firstly, as can be seen from Table 5, for LRR, the values of TPR and FPR are the highest on each data. This shows that the LRR method is sensitive to the extremely imbalanced datasets when learning

Table 4: The ratios of two types of samples on each multigenome data.

Multigenome data	n_T	n_N	n_T/n_N	n_N/n_T
COInteg	262	19	13.79	0.07
ESInteg	183	9	20.33	0.05
HNInteg	398	20	19.90	0.05

Table 5: The true positive rate and false positive rate on multigenome data.

Multigenome data	Metrics	*K*-means	GNMF	gLPCA	LRR	LLRR	BLLRR
COInteg	TPR	88.80	88.79	98.15	**100.00**	98.89	99.92
	FPR	38.00	100	48.84	94.74	67.19	**24.21**
ESInteg	TPR	99.45	99.45	98.61	**100.00**	99.11	99.45
	FPR	66.67	66.67	78.22	88.89	66.30	**55.56**
HNInteg	TPR	85.03	88.21	93.37	**99.75**	99.46	99.44
	FPR	71.00	79.00	86.00	100.00	67.00	**34.33**

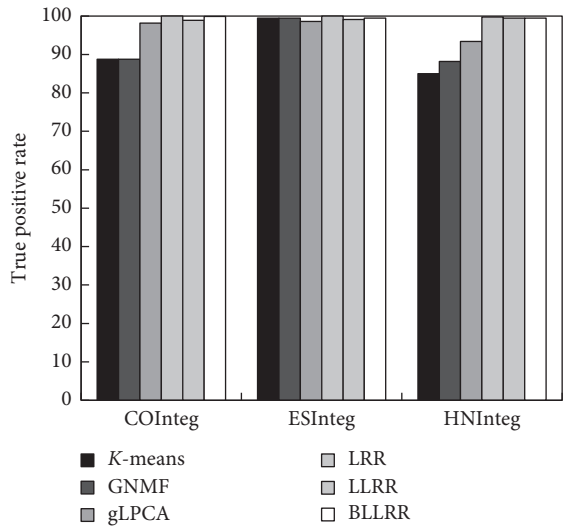

Figure 3: The true positive rates of six methods on each multigenome data.

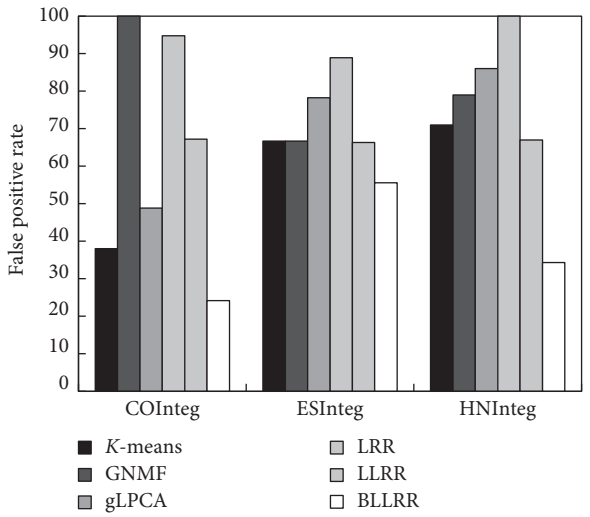

Figure 4: The false positive rates of six methods on each multigenome data.

subspace. That is, when the dataset is extremely unbalanced, the LRR method can only learn the subspace structure of majority class samples well but cannot learn the subspace structure of minority class samples well. So, LRR is not suitable for the study of subspace clustering in the case of extremely unbalanced samples. Secondly, as can be seen from Figure 4, compared with LRR, LLRR improves the clustering performance of minority class samples. This further shows that graph regularization helps to learn subspace information better by preserving local geometric structures in high-dimensional data, which is of great significance for the clustering of minority class samples. Finally, we compare BLLRR with LLRR. We can see from Figure 4 that, on each data, the FPR value of the BLLRR method is far less than that of the LLRR method and is the smallest of all the comparison methods. This shows that block-constraint is beneficial to extract more abundant structural information from multigenome data, thus avoiding the loss of the intrinsic subspace structure of minority class samples in manifold learning. In addition, this experimental result also proves the validity of the BLLRR method for clustering samples on extremely unbalanced data. To sum up, BLLRR can effectively learn the subspace structure embedded in multigenome data so that BLLRR can still cluster each class of samples effectively even though the samples are extremely unbalanced.

4. Conclusion

In this paper, we put forward a novel method termed BLLRR to analyze integrated TCGA data. In the BLLRR model, the graph Laplacian is introduced to make the BLLRR method respect the local geometric relationship of data better when learning the manifold structure. In addition, in order to deal with heterogeneous data better, the idea of block-constraint is introduced, which makes it convenient for BLLRR to impose different constraint intensities on different data

blocks. Because block-constraint can well balance the complex noise of multiclass data and better preserve the useful characteristic information of each class of data, our method is competent to learn the subspace structure of multiple heterogeneous data. Then, we apply the BLLRR method to cancer sample clustering based on multigenome data. Firstly, the integrated multigenome data are decomposed by BLLRR, and a coefficient matrix is obtained. Secondly, we construct the affinity matrix to denote the affinities between samples based on the coefficient matrix. Finally, we regard sample clustering as a problem of graph segmentation and use K-means to achieve the cancer sample clustering. The experimental results show that our method has remarkable subspace learning ability. Especially for minority class samples in extremely unbalanced datasets, the clustering performance of the BLLRR method is obviously better than other methods. So, the BLLRR method is an efficient and reliable method for multigenome data analysis. In future, we will continue to work on the comprehensive analysis of TCGA data.

Data Availability

The Data sets supporting the findings of this work are available at https://cancergenome.nih.gov/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the NSFC under grant nos. 61872220 and 61702299.

References

[1] K. A. Cronin, A. J. Lake, S. Scott et al., "Annual report to the nation on the status of cancer, part I: national cancer statistics," *Cancer*, vol. 124, no. 13, pp. 2785–2800, 2018.

[2] What is cancer?, https://www.cancer.gov/about-cancer/understanding/what-is-cancer-2019.

[3] R. P. Woychik, M. L. Klebig, M. J. Justice, T. R. Magnuson, and E. D. Avrer, "Functional genomics in the post-genome era," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 400, no. 1-2, pp. 3–14, 1998.

[4] C. K. Sarmah and S. Samarasinghe, "Microarray gene expression: a study of between-platform association of Affymetrix and cDNA arrays," *Computers in Biology and Medicine*, vol. 41, no. 10, pp. 980–986, 2011.

[5] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annual Review of Biomedical Engineering*, vol. 4, no. 1, pp. 129–153, 2002.

[6] A. Mortazavi, B. A. Williams, K. Mccue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628.

[7] The Cancer Genome Atlas Program, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga-2019.

[8] Z. Yu, H. S. Wong, and H. Wang, *Graph Based Consensus Clustering for Class Discovery from Gene Expression Data*, Oxford University Press, Oxford, UK, 2007.

[9] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[10] J. J.-Y. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer;clustering," *BMC Bioinformatics*, vol. 14, no. 1, p. 107, 2013.

[11] X.-Z. Kong, J.-X. Liu, C.-H. Zheng, M.-X. Hou, and J. Wang, "Robust and efficient biomolecular clustering of tumor based on-norm singular value decomposition," *IEEE Transactions on Nanobioscience*, vol. 16, no. 5, pp. 341–348, 2017.

[12] C.-M. Feng, Y.-L. Gao, J.-X. Liu, C.-H. Zheng, and J. Yu, "PCA based on graph laplacian regularization and P-norm for gene selection and clustering," *IEEE Transactions on Nanobioscience*, vol. 16, no. 4, pp. 257–265, 2017.

[13] A. K. Virmani, J. A. Tsou, K. D. Siegmund et al., "Hierarchical clustering of lung cancer cell lines using DNA methylation markers," *Cancer Epidemiology, Biomarkers & Prevention: a Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 11, no. 3, pp. 291–297, 2002.

[14] F. Ye, J. F. Xia, Y. W. Chong, Y. Zhang, and C. H. Zheng, "Tumor clustering using independent component analysis and adaptive affinity propagation," in *Proceedings of the International Conference on Intelligent Computing*, pp. 34–40, Taiyuan, China, August 2014.

[15] J.-X. Liu, Y. Xu, C.-H. Zheng, H. Kong, and Z.-H. Lai, "RPCA-based tumor classification using gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 964–970, 2015.

[16] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression," *BMC Bioinformatics*, vol. 15, no. 1, p. 37, 2014.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.

[18] Q. Zhang, J. E. Burdette, and J. P. Wang, "Integrative network analysis of TCGA data for ovarian cancer," *Bmc Systems Biology*, vol. 8, no. 1, p. 1338, 2014.

[19] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[20] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, pp. 597–610, 2013.

[21] H. Chang, "Learning discriminative low-rank representation for image classification," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 313–318, Beijing, China, July 2014.

[22] G. Liu, H. Xu, and S. Yan, "Exact subspace segmentation and outlier detection by low-rank representation," *Mathematics*, vol. 16, pp. 409–421, 2014.

[23] Q. Qu, N. M. Nasrabadi, and T. D. Tran, "Abundance estimation for bilinear mixture models via joint sparse and low-rank representation," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 52, pp. 4404–4423, 2014.

[24] J. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 8, pp. 1432–1445, 2014.

[25] N. Zhang and J. Yang, "Low-rank representation based discriminative projection for robust feature extraction," *Neurocomputing*, vol. 111, pp. 13–20, 2013.

[26] J. Feng, Z. Lin, H. Xu, and S. Yan, "Robust subspace segmentation with block-diagonal prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3818–3825, Columbus, OH, USA, June 2014.

[27] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 33, pp. 1548–1560, 2011.

[28] M. Zheng, J. Bu, C. Chen et al., "Graph regularized sparse coding for image representation," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1327–1336, 2011.

[29] X. Long, H. Lu, Y. Peng, and W. Li, "Graph regularized discriminative non-negative matrix factorization for face recognition," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2679–2699, 2014.

[30] J. Y. Wang, I. Almasri, and X. Gao, "Adaptive graph regularized Nonnegative Matrix Factorization via feature selection," in *Proceedings of the International Conference on Pattern Recognition*, pp. 963–966, Tsukuba, Japan, November 2012.

[31] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, "Learning Hypergraph-Regularized Attribute Predictors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 25, pp. 409–417, Boston, MA, USA, June 2015.

[32] Y. Peng, S. Wang, X. Long, and B.-L. Lu, "Discriminative graph regularized extreme learning machine and its application to face recognition," *Neurocomputing*, vol. 149, pp. 340–353, 2015.

[33] W. Cheng, X. Zhang, Z. Guo, Y. Shi, and W. Wang, "Graph-regularized dual Lasso for robust eQTL mapping," *Bioinformatics*, vol. 30, no. 12, pp. i139–i148, 2014.

[34] Z. Wang, Q. Ruan, G. An, and Y. Jin, "A regularized low-rank representation model for facial expression recognition," in *Proceedings of the IEEE International Conference on Signal Processing*, pp. 1072–1076, London, UK, August 2017.

[35] X. Lu, Y. Wang, and Y. Yuan, "Graph-regularized low-rank representation for destriping of hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 4009–4018, 2013.

[36] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 504–517, 2016.

[37] Y.-X. Wang, J.-X. Liu, Y.-L. Gao, C.-H. Zheng, and J.-L. Shang, "Differentially expressed genes selection via Laplacian regularized low-rank representation method," *Computational Biology and Chemistry*, vol. 65, pp. 185–192, 2016.

[38] S. Du, Y. Ma, and Y. Ma, "Graph regularized compact low rank representation for subspace clustering," *Knowledge-Based Systems*, vol. 118, pp. 56–69, 2016.

[39] B. Gan, C. H. Zheng, J. Zhang, and H. Q. Wang, "Sparse representation for tumor classification based on feature extraction using latent low-rank representation," *BioMed Research International*, vol. 2014, Article ID 420856, 7 pages, 2014.

[40] J. Wang, J. X. Liu, C. H. Zheng, Y. X. Wang, X. Z. Kong, and C. G. Weng, "A mixed-norm laplacian regularized low-rank representation method for tumor samples clustering," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 16, no. 1, pp. 172–182, 2017.

[41] C. Q. Xia, K. Han, Y. Qi, Y. Zhang, and D. J. Yu, "A self-training subspace clustering algorithm under low-rank representation for cancer classification on gene expression data," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 15, no. 4, pp. 1315–1324, 2017.

[42] J. Wang, J.-X. Liu, X.-Z. Kong, S.-S. Yuan, and L.-Y. Dai, "Laplacian regularized low-rank representation for cancer samples clustering," *Computational Biology and Chemistry*, vol. 78, pp. 504–509, 2018.

[43] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1735–1742, New York, NY, USA, June 2006.

[44] M. Belkin and P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, MIT Press, Cambridge, MA, USA, 2003.

[45] R. K. C. Fan, *Spectral Graph Theory*, American Mathematical Society, Providence, RI, USA, 1997.

[46] M. Tao and X. Yuan, "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *Siam Journal on Optimization*, vol. 21, no. 1, pp. 57–81, 2011.

[47] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Mathematics of Computation*, vol. 82, no. 281, pp. 301–329, 2012.

[48] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," *Advances in Neural Information Processing Systems*, pp. 612–620, Cornell University, Ithaca, NY, USA, 2011.

[49] J. F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *Siam Journal on Optimization*, vol. 20, pp. 1956–1982, 2008.

[50] G. Tang and A. Nehorai, "Robust principal component analysis based on low-rank and block-sparse matrix decomposition," in *Proceedings of the 45th Annual Conference on Information Sciences and Systems*, Princeton, NJ, USA, May 2011.

[51] J. R. Pollack, C. M. Perou, Aa et al., "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genetics*, vol. 23, no. 1, pp. 41–46, 1999.

[52] P. Marttinen, S. Myllykangas, and J. Corander, "Bayesian clustering and feature selection for cancer tissue samples," *BMC Bioinformatics*, vol. 10, no. 1, p. 90, 2009.

[53] F. Watt and P. L. Molloy, "Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter," *Genes & Development*, vol. 2, no. 9, pp. 1136–1143, 1988.

[54] J. Boyes and A. Bird, "DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein," *Cell*, vol. 64, no. 6, pp. 1123–1134, 1991.

[55] J. Bo, C. Ding, B. Luo, and T. Jin, "Graph-laplacian PCA: closed-form solution and robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OE, USA, June 2013.

[56] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–273, Toronto, Canada, August 2003.

[57] L. Lovász and M. D. Plummer, "Matching theory," *Annals of Discrete Mathematics 1986*, vol. 29, 1986.