Journal of Medical Diagnostic Methods
ISSN: 2168-9784

Kim, et al., J Med Diagn Meth 2015, 4:4
DOI; 10.4172/2168-9784.1000189

Research Article Open Access

Identification of Tumor Subtypes of Endometrial Carcinoma by Integration of Heterogeneous Datasets

Hyunsoo Kim[1*], Markus Bredel[2], Haesun Park[3] and Jeffrey H. Chuang[1]

[1]*The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA*
[2]*Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, AL, USA*
[3]*School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA*

Abstract

The Cancer Genome Atlas (TCGA) project has made available multiple heterogeneous datasets. Although several methodological approaches have been proposed for the heterogeneous data integration, there is no framework of sparse non-negative matrix factorization (NMF) for handling heterogeneous biological data integration. Here, we propose the block-weighted sparse NMF (bwsNMF) to identify tumor subtypes of endometrial carcinoma by integrating gene expression, mutations, a protein-protein interaction network and a transcription factor target network.

Keywords: Cancer subtype; Personalized medicine; Weighted sparse NMF

Introduction

Clustering algorithms have been applied to the identification of new subtypes of human cancer. Clustering of heterogeneous datasets represents a difficult clustering problem to which some clustering methods cannot be easily extended. Clustering methods based on matrix computations, such as non-negative matrix factorization (NMF), can be modified to deal with this complex problem. In this paper, we will show how the formulation of NMF can be modified for tumor subtype identification with multiple heterogeneous datasets.

An important question concerns the application of different weights to the multiple heterogeneous datasets. A gold standard reference is needed to choose the weight parameters. In supervised classification problems or semi-supervised clustering problems, we have a training set that can be used for parameter-tuning with cross-validation. However, our objective is to determine weight parameters when there is no training set that has subtype labels. We use another type of data, i.e. clinical data including survival days of patients, to determine weight parameters after defining the best subtypes consisting of patient groups that show different survival profiles. In other words, our objective is to identify tumor subtypes that maximize survival differences by searching the weight parameter space. We believe that NMF provides a useful mathematical framework to formulate a more complex objective function without losing computational efficiency.

The current general approach to personalized cancer genomic medicine is based on the identification of cancer subtypes from genomic profiles. The most widely used genomic profile is derived from gene expression data. Because some cancer types are driven by somatic mutations or copy number aberrations, it is logical to use multiple heterogeneous datasets to identify more meaningful cancer subtypes with clinical relevance.

Tumor subtyping is a classic clustering problem that is one of the major focus areas in computer science and statistics. Although there are many clustering algorithms being developed, it is difficult to assess which algorithm is the best performer, because they tend to be situation-specific, and the performance of each algorithm generally depends on the dataset. Nevertheless, there is a strong rationale to choose an algorithm that generally performs well to identify clusters in a given dataset and allows for easier interpretation, improving our ability to understand the given data and identify new biological knowledge. This is one of the reasons why many computational biologists have used hierarchical clustering (HC) with various distance metrics. One of common sub-problems with this approach is how we can determine the number of clusters. In order to identify the number of clusters and membership stability, consensus HC was developed [1] in 2003. Some authors who developed the consensus HC later participated in applying an NMF algorithm [2] to bioinformatics and computational biology to discover metagenes and molecular patterns [3]. However, the NMF algorithm based on multiplicative update rules (i.e. gradient descent method) can suffer from convergence issue [4]. Therefore, some NMF algorithms [5,6] based on Newton's method have been developed and applied to bioinformatics and computational biology [7]. These NMF algorithms based on alternating least squares usually showed faster convergence speed, so they have been implemented by multiple computer languages (e.g., NMF Matlab toolbox [8] in Matlab, NIMFA in Python [9]). Recently, subtypes of adult de novo acute myeloid leukemia have been identified by NMF instead of HC [10].

The Cancer Genome Atlas (TCGA) project has generated multi-type genomic datasets as well as clinical data. Tumor subtyping has been primarily done with gene expression profiles, because gene expression is considered to be the output of all regulatory variations/profiles such as SNPs, mutations, structural variations (SVs), copy number variations (CNVs), microRNA profiles and DNA methylation profiles. The TCGA community has proposed four subtypes (classical, neural, mesenchymal and proneural) in glioblastoma [11] with gene expression profiles and four main breast cancer subtypes (cluster1: similar to PAM50 HER2-enriched, cluster2: PAM50 Basal, cluster3:

***Corresponding author:** Hyunsoo Kim, The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA, Tel: +18608562495; Fax: +18608562398; E-mail: hyunsoo.kim@jax.org

Received November 25, 2015; **Accepted** December 18, 2015; **Published** December 18, 2015

Citation: Kim H, Bredel M, Park H, Chuang JH (2015) Identification of Tumor Subtypes of Endometrial Carcinoma by Integration of Heterogeneous Datasets. J Med Diagn Meth 4: 189. doi: 10.4172/2168-9784.1000189

PAM50 Luminal A, and cluster4: PAM50 Luminal B) when combining five heterogeneous datasets (miRNA, DNA methylation, copy number, gene expression and reverse-phase protein array (RPPA) dataset) [12] with consensus HC. mRNA expression profiles of colorectal cancer were clustered into three distinct clusters [13,14], but they were not associated with any clinical phenotype such as patient survival or response to chemotherapy [15]. mRNA expression profiles of lung squamous cell carcinoma (SQCC) were categorized into four subtypes (classical, basal, secretory and primitive) [16] with the previously reported lung SQCC gene expression-subtype signatures.

In this paper, we propose the block-weighted sparse NMF (bwsNMF) to identify subtypes of endometrial carcinoma by integrating gene expression data, gene mutations, a protein-protein interaction network, and a transcription factor target network.

Material and Methods

Sparse NMF

We briefly review a type of sparse NMF (sNMF) based on alternating non-negativity-constrained least squares [6] to impose sparseness constrained on basis/metagene matrix:

$$\min_{W,H\geq 0} A-\| WH\|_F^2 \|+\eta \| H\|_F^2\|+\alpha\sum_{i=1}^{m}\| W(i,:)\|_1^2, \quad (1)$$

where $A\in\mathbb{R}_+^{m\times n}$ is the non-negative input (m genes × n patients) matrix, $W\in\mathbb{R}_+^{m\times k}$ is the basis/metagene matrix, $H\in\mathbb{R}_+^{k\times n}$ is the coefficient/loading matrix, $W(i,:)$ is the ith row vector of W, η>0 is a parameter to suppress $\|H\|_F^2$ and $\|\cdot\|_1$ is a regularization parameter to balance the trade-off between accuracy of approximation the sparseness of W [6]. Because this formulation has the L1-norm (i.e. $\|.\|_1$) term, it is theoretically appropriate to introduce sparseness (exact zero values in W). Moreover, the formulation can be optimized by least squares methods due to non-negativity-constraints, which results in high convergence speed with sound convergence property. The L1-norm term can be considered without using linear programming. This particular type of sNMF is different from Lasso (a shrinkage and selection method for linear regression) [17], because Lasso has no square in the L1-norm term. Although sNMF [6] for sparser $H\in R_+^{k\times n}$ has been applied to clustering problems, sNMF for sparser $W\in R_+^{m\times k}$ has not been applied to a clustering problem yet.

We can expand the term of

$$\alpha\sum_{i=1}^{m}\| W(i,:)\|_1^2$$

$$\alpha\sum_{i=1}^{m}\| W(i,:)\|_1^2=\alpha\left(\left(w_{11}+\ldots+w_{1k}\right)^2+\left(w_{21}+\ldots+w_{2k}\right)^2+\ldots+\left(w_{m1}+\ldots+w_{mk}\right)^2\right)$$

where $w_{ij}\geq 0$ is the ith row and the jth column of W≥0, k is the number of basis vectors, i.e. the number of clusters, thus, the objective function can be optimized by alternating non-negativity-constrained least squares:

$$\min_{W\geq 0}\left\|\begin{pmatrix}H^T\\ \sqrt{\alpha}\mathbf{e}_{1\times k}\end{pmatrix}W^T-\begin{pmatrix}A^T\\ 0_{1\times m}\end{pmatrix}\right\|_F^2 \quad (2)$$

where $\mathbf{e}_{1\times k}\in\mathbb{R}^{1\times k}$ is an all-ones-vector and $0_{1\times m}\in\mathbb{R}^{1\times m}$ is a zero vector, and

$$\min_{H\geq 0}\left\|\begin{pmatrix}W\\ \sqrt{\eta}\mathrm{I}_{k\times k}\end{pmatrix}H-\begin{pmatrix}A\\ \mathrm{O}_{k\times n}\end{pmatrix}\right\|_F^2 \quad (3)$$

where $\mathrm{I}_{k\times k}\in\mathbb{R}^{k\times k}$ is an identity matrix and $\mathrm{O}_{k\times n}\in\mathbb{R}^{k\times n}$ is a zero matrix [6]. This proposed algorithm is a two-block block coordinate descent (BCD) method, which satisfies Bertsekas conditions [18] for convergence (i.e. each sub-problem has a unique solution). By imposing sparseness on the basis/metagene matrix, contribution of some genes to each basis vector can be converted to zero, which may give us simpler basis vectors that consist of more important genes/features to describe the original input matrix.

Block Weighted Sparse NMF

Here, we introduce a novel NMF called the block-weighted sparse NMF (bwsNMF) that can consider the relative importance of feature blocks as well as sparseness. When we know a set of features may have a higher or lower priority than other sets, we can divide $A\in\mathbb{R}_+^{m\times n}$ into multiple row blocks: $A=[A_1;A_2;\ldots;A_d]$ where $1\leq l\leq d$ is the lth block matrix (m_l genes × n patients, $1\leq l\leq d$ where d is the total number blocks) of $A\in\mathbb{R}_+^{m\times n}$ Then, the objective function of bwsNMF follows:

$$\min_{W_1,W_2,\ldots,W_d,H\geq 0}\| A_1-W_1H\|_F^2+\ldots+\| A_d-W_dH\|_F^2+\eta\| H\|_F^2+\sum_{l=1}^{d}\left(s_l\sum_{i=1}^{m_l}\| W_l(i,:)\|_1^2\right), \quad (4)$$

Where $W_l\in\mathbb{R}_+^{m_l\times k}$ is the lth block matrix of the basis/metagene matrix $W\in\mathbb{R}_+^{m\times k}$ H is the coefficient/loading matrix $W_l(i,:)$ is the ith row vector of $W_l\in\mathbb{R}_+^{m_l\times k}$, $\eta>0$ is a parameter to suppress $\|H\|_F^2$ and s_l > 0 is a regularization parameter to introduce zeros into W_l. We can still optimize the objective function with alternating non-negativity-constrained least squares:

$$\min_{W_1\geq 0}\left\|\begin{pmatrix}H^T\\ \sqrt{s_1}\mathbf{e}_{1\times k}\end{pmatrix}W_1^T-\begin{pmatrix}A_1^T\\ 0_{1\times m}\end{pmatrix}\right\|_F^2 \quad (5)$$

$$\min_{W_2\geq 0}\left\|\begin{pmatrix}H^T\\ \sqrt{s_2}\mathbf{e}_{1\times k}\end{pmatrix}W_2^T-\begin{pmatrix}A_2^T\\ 0_{1\times m}\end{pmatrix}\right\|_F^2 \quad (6)$$

$$\min_{W_d\geq 0}\left\|\begin{pmatrix}H^T\\ \sqrt{s_d}\mathbf{e}_{1\times k}\end{pmatrix}W_d^T-\begin{pmatrix}A_d^T\\ 0_{1\times m}\end{pmatrix}\right\|_F^2 \quad (7)$$

Where $\mathbf{e}_{1\times k}\in\mathbb{R}^{1\times k}$ is an all-ones-vector, $0_{1\times m}\in\mathbb{R}^{1\times m}$ is a zero vector. The next alternating step is

$$\min_{H\geq 0}\left\|\begin{pmatrix}W\\ \sqrt{\eta}\mathrm{I}_{k\times k}\end{pmatrix}H-\begin{pmatrix}A\\ \mathrm{O}_{k\times n}\end{pmatrix}\right\|_F^2 \quad (8)$$

where $W\in\mathbb{R}_+^{m\times k}$ is the merged matrix from $[W_1;W_2;\ldots;W_l]$ with preserved order, $\mathrm{O}_{k\times n}\in\mathbb{R}^{k\times n}$ is an identity matrix, and $\mathrm{O}_{k\times n}\in\mathbb{R}^{k\times n}$ is a zero matrix. The convergence criteria of sNMF can be naturally applied to this algorithm, and the fast convergence property is also preserved. This bwsNMF can incorporate biological knowledge with the weight values (s_1,s_2,...,s_l) Larger sl gives higher chance to introduce zeros to $W_l\in\mathbb{R}_+^{m_l\times k}$ with a set of specific features for the lth block.

When we know the relative importance of feature blocks, we can reduce the searching space of the weight parameters. When we have a biological hypothesis that one feature type is more important than the other feature type for separating a clinical phenotype, we can set higher priority to the more important feature type, obtain tumor subtypes, and measure the degree of associations between subtypes and clinical phenotypes so that we can assess the quality of the biological hypothesis. When we do not have any a priori knowledge and assumptions regarding the relative importance of a feature type, we search for a set of parameters that can generate maximum separation of a clinical phenotype (e.g., survival) of patients in different groups.

Tumor Subtype Identification with bwsNMF

Tumor subtypes have been determined by consensus HC1 or consensus NMF3, where the corresponding clustering algorithms were applied to genes usually selected by removing genes with small variances and/or removing gene profiles with low absolute expression values. Thus, the typical number of genes in the input matrix $A \in \mathbb{R}^{m \times n}$ is less than 10,000 genes (i.e. m<10,000), which is a relatively small number compared to the number of total genes in the human genome. This gene selection step can be considered as a statistical treatment rather than the usage of biological knowledge.

As an example to show bwsNMF can incorporate multi-type biological knowledge, we tried to identify tumor subtypes of endometrial carcinoma with the following features: 1) somatic mutations, 2) neighbor genes of the somatic mutations, and 3) target genes of transcription factors directly or indirectly connected with somatic mutations in the cancer pathways (called as TF-target genes). Then, the typical number of genes in the input matrix $A \in \mathbb{R}^{m \times n}$ was reduced to fewer than 1000 genes (i.e. m<1,000)). High quality biological databases can be used for this initial feature selection step. However, a particular challenge is predicting the relative importance of multi-type features. Therefore, bwsNMF was repeatedly executed with a set of weight parameters inside the parameter searching space (10-8, 10-6…,1) for each sl, to find the best set of parameters that can identify tumor subtypes for which patient survival was maximally different.

Results

The datasets generated by TCGA are unique in that they have multi-level genomic profiles for the same patients and well-controlled clinical data including survival, drug response and histological characterization. We decided to use TCGA data to assess the effectiveness of our bwsNMF algorithm in regard to tumor subtype identification. We downloaded mutation data and RNA-Seq data of uterine corpus endometrial carcinoma from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) in August, 2013. The RNA-Seq data obtained from Illumina Genome Analyzer with RNASeqV2 protocol contained more than 20,000 genes and 370 tumor samples. We applied sNMF and bwsNMF for various input matrices built by different coding schemes to integrate different data types with η=1.

sNMF with Highly Varied Genes in RNA-Seq

We built a large biological network with published molecular interaction networks [19,20] to incorporate biological knowledge for tumor subtype identification. We tested a mutation indicator input matrix, of which each element was a fixed number for mutated genes and their neighbor genes (maximal depth=5 in the protein–protein interaction network), or a zero value for otherwise in a corresponding sample. This coding scheme was similar to, but still different from that of the network-based stratification (NBS) [15]. We did not apply any normalization to the indicator matrix. We used 194 mutated genes for which mutations occur in at least 50 times, which resulted in 833 genes (mutated genes + their neighbour genes). sNMF was repeatedly applied to the dataset with different α values to obtain the most well-separated survival curves among groups. In this case, three groups (group1: 208, group2: 134, group3: 28) were suggested, but they were not associated with overall survival (log-rank P=0.150). This could have several explanations. Some of patients did not have these mutations. This scheme does not consider the concentration of mutated genes or their neighbour genes. When some mutations observed in DNA exome sequences are very lowly expressed across patients, they might not generate any effect on the downstream pathways and not produce any power to discriminate patient groups. Some gain-of-function mutations became effective only by interaction with an important modifier-gene. When the modifier-genes were very lowly expressed across patients, the effect of the gain-of-function mutations was limited.

We tested another mutation indicator input matrix by selecting only 115 highly varying genes in the RNA-Seq profiles among 833 genes selected by the above scheme (i.e. frequently mutated genes and their neighbour genes). The first group (243 patients) usually did not have mutations of these genes or their neighbours. The second group (73 patients) had mutation neighbour genes (ENPP3, FOXJ1, LY6D, PRAP1 and SLCO2A1). Note that mutation neighbour genes instead of mutations determine this second group. The third group (54 patients) had many mutations and their neighbour genes in a protein–protein network, which was driven by mutations. The degree of association with overall survival was slightly better than the previous case (P=0.107). This observation supports the notion that it is meaningful to select mutations and their neighbour genes of which gene expression profiles were varied across samples because mRNA concentration of mutation genes or their binding partner may affect tumor progression and/or metastasis. However, using only mutations was not sufficient to inform overall survival.

bwsNMF With RNA-Seq, Mutation Information, and Biological Networks

Because we hypothesized that including transcription factor target genes was important for clustering as well, we decided to include them and applied different weights on them with bwsNMF, because we had two different blocks and we did not know which block was more important for clustering. bwsNMF was used to cluster an input matrix with 115 genes highly varied in RNA-Seq profiles selected from 833 mutated genes and their nearest neighbor genes whose distance to a mutation was less than six in the protein-protein interaction network, and highly varied 235 target genes [20] of five transcription factors (MYC, CTCF, ESR1, ATF3, TP53). Thus, the total number of selected genes was 350. The input matrix has RNA-Seq expression levels for the transcription factor target genes and a fixed value (e.g., the largest RNA-Seq expression level) for mutations and their neighbors (called as mutation-related genes). When a gene was a transcription factor target gene and also a mutation-related gene, we used the fixed value to indicate that it was a mutation-related gene, although it was also a target gene. Interestingly, bwsNMF with this input matrix-coding scheme identified three novel subgroups highly associated with overall survival (log-rank P=0.009) (Figure 1). The number of patients for each subgroup was 206, 143 and 21. Figure 2 shows the basis/metagene matrix and the coefficient matrix obtained from bwsNMF. We only drew the discriminative genes that were differentially expressed between tumor types in the metagene matrix. The discriminative genes of group 1 were C4BPA, EDN3, PGR, SERPINA3 and SLC47A1, which were targets of ESR1. CBS, IGF2BP1,

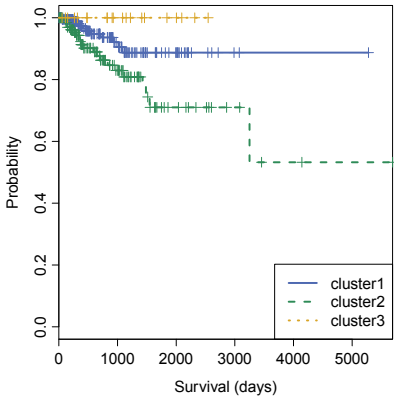

Figure 1: Kaplan-Meier curves of overall survival for three clusters identified by block weighted sparse NMF with an input matrix of which elements have an indicator value for mutated genes and their nearest neighbors whose distance to a mutated gene was less than six in the protein-protein interaction network, and RNA-Seq gene expression values for highly varied transcription factor target genes.

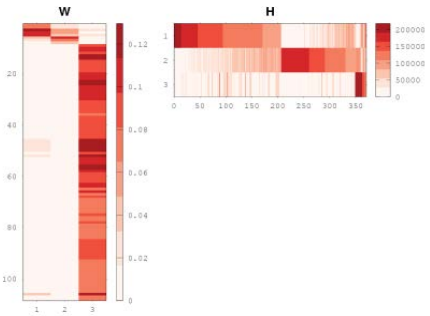

Figure 2: The metragene matrix (left) and the coefficient matrix (right) obtained from our block weighted sparse NMF. The metagene matrix was only drawn with discrimitive genes that were differently expressed between tumor subtypes. Each column of the metagene matrix is a basis vector. Each column of the coefficient matrix is a three-dimensional representation of a patient's genomic profile.

and FCHO1 were highly expressed in the second group, where CBS and IGF2BP1 were targets of MYC. Mutations and their neighbour were enriched in group 3. Each column of the coefficient matrix indicates the three-dimensional representation of a patient's genomic profile. Figure 3 shows the sample-sample relationship with the three-dimensional representations of the samples. sNMF with the same input matrix revealed three groups associated with overall survival at a lower level of significance (log-rank P=0.01). Collectively, using biological knowledge (mutated nodes, mutation neighbour nodes in a protein-protein interaction network and a transcription target network) in combination with bwsNMF was very powerful in identifying new, clinically relevant subgroups of endometrial carcinoma.

Discussion

Here, we discuss the difference between our block-weighted sparse NMF and a graph regularized NMF-based approach. The network-based

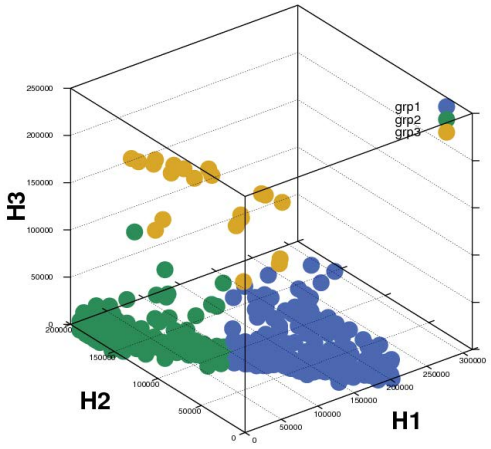

Figure 3: The coefficient matrix obtained from our block weighted sparse NMF has the three-dimensional representations of samples. Each row of the coefficient matrix means the three-dimensional coordinate of this figure. H1, H2, and H3 are corresponding to the first, the second, and the third axis, respectively.

stratification (NBS) [15] used graph regularized NMF (GNMF) [21]

$$\min_{W,H\geq 0} \| A-WH \|_F^2 + \lambda \operatorname{Tr}\left(W^T L W\right), \quad (9)$$

where A is the non-negative input (gene X patient) matrix, W is the basis/metagene matrix, H is the coefficient/loading matrix, L=D-K is graph Laplacian [22], where D is a diagonal matrix whose entries are column sums of K $D_{ii} = \sum_j K_{ij}$ and K_{ij} is the weight element of the ith row and the jth column, and Tr (·)denotes the trace of a matrix. The term of Tr (W^T LW) constrains the column vectors of W to give higher weight to local graph neighborhoods to take into account local graph topology around each mutated somatic gene. NBS used mutation information for tumour subtype identification. Its performance depends not only on network smoothing but also on NMF, because its performance was improved with NMF instead of HC [15]. NBS took advantage of somatic mutation information and a gene–gene interaction network by projecting mutations onto the network for the clustering of ovarian serous cystadenocarcinoma into four subtypes, lung adenocarcinoma into six subtypes, and endometrial carcinoma into three subtypes that are associated with patient overall survival or histological type. The most predictive data types were somatic mutations for ovarian cancer patient survival, somatic mutations and RNA-Seq for lung cancer patient survival, and copy number variations for endometrial cancer patient histological types [15], respectively. The minimization process of the objective function was similar to Lee and Seung's multiplicative updating algorithm [23] that suffers from convergence issues. The network-regularized NMF code for NBS was extended from the NMF Matlab toolbox [8] partly to consider convergence issues, because this particular implementation was based on multiplicative update rules to obtain W and the non-negativity-constrained least squares or multiplicative update rules to obtain H. This implementation was helpful for handling convergence issues, but it still used multiplicative update rules at the one-side of alternating steps at least. In order to

completely resolve the convergence issues, we need to find a way to avoid the multiplicative update rules. By contrast, our block weighted sparse NMF is based on the block coordinate descent method that generally shows faster convergence than multiplicative updating rules. The NBS approach does not have a way to apply different weights to different data types, and it does not use a transcriptional regulatory network. Our bwsNMF algorithm is an advanced form of sparse NMF, which is designed to apply different weights to heterogeneous data types, and our strategy for tumour subtype identification uses a transcription regulatory network, because target genes of transcription factors that are downstream effectors of somatic mutations may represent distinct biological subtypes.

We reviewed the published survival analyses to assess the significance of endometrial cancer subtypes identified by various methods [15,24] from TCGA datasets. In the TCGA paper on endometrial carcinoma [24], somatic copy number aberrations (SCNAs) were hierarchically clustered into four groups associated with progression-free survival (log-rank P=0.0004) [24]. Multiple profiles were used to compare the four groups (POLE ultramutated, microsatellite instability (MSI) hypermutated, copy-number low, and copy-number high) associated with progression-free survival (log-rank P=0.02) [24]. The five clusters defined by RPPA profiles were highly correlated with histology (endometrioids, mixed, serous), grade and clusters defined by other platforms (mRNA three groups (mitotic, hormonal, immunoreactive), CNAs, DNA methylation four groups, microRNA, and MHL1 hypermethylation) [24]. MicroRNA profiles were clustered into six groups associated with overall survival (log-rank P=0.18) [24]. iCluster revealed two groups (P=0.077 for overall survival, P=0.23 for recurrence-free survival) from somatic mutation, DNA copy number, DNA methylation, and mRNA expression data [24]. SuperCluster [24] identified four groups (hyper-mutator super cluster, low mutator endometrioid super cluster, ultra-mutator super cluster, serous super cluster), which showed association with cancer specific overall survival (P=0.0367) and progression-free survival (P=0.0265) [24]. However, any strong association between three groups identified from mRNA-Seq and patient overall survival or progression-free survival has not been reported. Although NBS unveiled three subgroups associated with histological types of endometrial carcinoma [15], association with overall survival or progression-free survival was not reported.

In summary, we applied sNMF for sparser $W \in \mathbb{R}_{+}^{m \times k}$ to a clustering problem and showed that it could identify tumor subtypes with clearer basis/metagene vectors by introducing more zeros in W. We developed a novel NMF algorithm of bwsNMF that could handle and assign different weights to heterogeneous datasets. We report three clinically relevant subtypes of the endometrial carcinoma, identified by bwsNMF by integration of mutation genes, mutation neighbor genes, and transcription factor target genes of mutation-related genes. The bwsNMF algorithm could be applied to other biomedical clustering problems with multiple heterogeneous data types.

References

1. Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning 52: 91-118.
2. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization Nature 401: 788-791.
3. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization Proc Natl Acad Sci U S A 101: 4164-4169.
4. Lin C-J (2007) On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. IEEE Transactions on Neural Networks 18: 1589-1596.
5. Kim H, Park H (2008) Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method. SIAM Journal on Matrix Analysis and Applications 30: 713-730.
6. Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis Bioinformatics 23: 1495-1502.
7. Kim H, Park H (2007) Cancer Class Discovery Using Non-negative Matrix Factorization Based on Alternating Non-negativity-Constrained Least Squares. In: Mandoiu I, Zelikovsky A, eds. Bioinformatics Research and Applications. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg 4463: 477-487.
8. Li Y, Ngom A (2013) The non-negative matrix factorization toolbox for biological data mining Source Code Biol Med 8: 10.
9. Zitnik M, Zupan B (2012) NIMFA? A Python Library for Nonnegative Matrix Factorization. Journal of Machine Learning Research 13: 849-853.
10. Cancer Genome Atlas Research Network (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia N Engl J Med 368: 2059-2074.
11. Verhaak RGW, Hoadley KA, Purdom E (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17: 98-110.
12. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours Nature 490: 61-70.
13. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, et al. (2010) Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types Clin Cancer Res 16: 4864-4875.
14. (2012) The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer Nature 487: 330-337.
15. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations Nat Methods 10: 1108-1115.
16. Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers Nature 489: 519-525.
17. Tibshirani R. (1994) Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B 58: 267–288.
18. Bertsekas DP (1999) Nonlinear Programming. (2nd Edtn). Belmont, Mass., USA: Athena Scientific
19. Costa PR, Acencio ML, Lemke N (2010) A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data BMC Genomics 11 Suppl 5: S9.
20. Bovolenta LA, Acencio ML, Lemke N (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions BMC Genomics 13: 405.
21. Cai D, He X, Han J, Huang TS (2011) Graph Regularized Nonnegative Matrix Factorization for Data Representation IEEE Trans Pattern Anal Mach Intell 33: 1548-1560.
22. Chung FRK. (1997) Spectral Graph Theory. American Mathematical Soc.
23. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization Nature 401: 788-791.
24. Kandoth C, Schultz N, Cherniack AD, Akbani R, Cancer Genome Atlas Research Network, et al. (2013) Integrated genomic characterization of endometrial carcinoma Nature 497: 67-73.