Molecular Pattern Discovery based on Penalized Matrix Decomposition

Chun-Hou Zheng, Lei Zhang, To-Yee Ng, Chi Keung Shiu, and De-Shuang Huang

Abstract: A reliable and precise identification of the type of tumors is crucial to the effective treatment of cancer. With the rapid development of microarray technologies, tumor clustering based on gene expression data is becoming a powerful approach to cancer class discovery. In this paper, we apply the penalized matrix decomposition (PMD) to gene expression data to extract metasamples for clustering. The extracted metasamples capture the inherent structures of samples belong to the same class. At the same time, the PMD factors of a sample over the metasamples can be used as its class indicator in return. Compared with the conventional methods such as hierarchical clustering (HC), self-organizing maps (SOM), affinity propagation (AP) and non-negative matrix factorization (NMF), the proposed method can identify the samples with complex classes. Moreover, the factor of PMD can be used as an index to determine the cluster number. The proposed method provides a reasonable explanation of the inconsistent classifications made by the conventional methods. In addition, it is able to discover the modules in gene expression data of conterminous developmental stages. Experiments on two representative problems show that the proposed PMD based method is very promising to discover biological phenotypes.

Index Terms: Tumor Clustering, Penalized Matrix Decomposition, Metasample, Gene Expression Data, Developmental Biology

C.-H. Zheng is with the College of Information and Communication Technology, Qufu Normal University, Rizhao, Shandong, China, and with the Biometric Research Center, Dept. of

Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: zhengch99@126.com).

L. Zhang is with the Biometric Research Center, Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (Corresponding author; phone: 852-27667355; e-mail: cslzhang@comp.polyu.edu.hk).

T.-Y. Ng is with the Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: cstyng@comp.polyu.edu.hk;).

C.-K. Shiu is with the Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: csckshiu@comp.polyu.edu.hk).

D.-S. Huang is with School of Electronics and Information Engineering, Tongji University, 4800 Caoan Road, Shanghai, China(e-mail: dshuang@iim.ac.cn).

1. Introduction

The rapid development of DNA microarray technology has made it possible to monitor gene expression levels on a genomic scale. The gene expression data captured using this high-throughput technique can potentially provide systematic information regarding to the underlying dynamics and mechanisms in biology, which enhances much the fundamental understanding of life on the molecular level. The challenge is how to interpret such data to gain insight into biological processes and the mechanisms of human diseases [12, 22, 25, 29, 36, 38]. Analysis of these data requires mathematical tools that are adaptable to the huge amount of data, and reducing the data complexity to make them comprehensible. Fortunately, many effective methods have been proposed for gene expression data analysis. Among them, clustering is a topical application to gene expression data for identifying the genes or samples with similar expression patterns [2, 3, 10, 23, 27, 37].

Many well-known clustering methods, such as hierarchical clustering (HC), self-organizing maps (SOM), affinity propagation (AP) and non-negative matrix factorization (NMF), have been successfully used for gene expression data clustering [5, 9, 10, 28, 30]. HC has been employed in analyzing temporal expression patterns [7], predicting patient outcomes among lymphoma patients [2], and providing molecular portraits of breast tumours [15]. However, one disadvantage of HC is that it imposes a stringent tree structure on the data. In addition, HC is highly sensitive to the metric used to assess similarity and often requires subjective evaluation to define clusters [5]. SOM provides another tool for clustering [20]. It has been successfully used to recognize the subtypes of leukemia [10]. SOM, however, is unstable. It yields different decompositions of the data depending on the choice of initial conditions [5]. Brunet et al. [5] demonstrated that NMF is more accurate than HC and it is more stable than SOM. Gao and George [9] showed that the results of NMF can be improved by using the sparse NMF (SNMF). Zheng et al. [28] improved the NMF clustering results by using gene selection methods.

Though these clustering algorithms are useful, one limitation of them is that each sample can only be clustered into one class, which may not be identical to the facts in some instances, e.g., borderline tumors and compound tumors [11, 14]. To overcome this problem, we propose to extract “metasamples” from the gene expression data by using Penalized Matrix Decomposition (PMD) [24]. A metasample is a linear combination of original samples. By using PMD to extract a small number of metasamples, each metasample can capture the inherent structures of the samples belong to the same class. At the same time, the samples can be clustered by mapping themselves to the extracted metasamples. Moreover, the number of metasamples, i.e., the number of clusters, could be determined according to the changing trend of factor

D (please refer to Eqs. (1) and (2)) extracted by PMD. The experiments show that the proposed method can identify the samples with complex classes, and it provides a reasonable explanation of the inconsistent classifications made by conventional methods such as HC, SOM, AP, spectral clustering (SC) [34] and NMF. Interestingly, it is also able to discover the modules in gene expression data of conterminous developmental stages. The contribution of this paper lies in the proposition of a PMD based clustering approach to molecular pattern discovery. It can detect compound tumors which cannot be discovered by conventional clustering methods.

The rest of the paper is organized as follows. Section 2 describes the methodology proposed in this study. Section 3 presents the numerical experiments. Section 4 concludes the paper and outlines directions of future work.

2. Methodology

2.1 Penalized Matrix Decomposition (PMD)

This subsection briefly introduces the PMD proposed by Witten et al. [24]. Consider a gene expression data set that consists of p genes in n samples. We denote it by a matrix X of size $p \times n$. Without loss of generality, we assume that the column and row means of X are zero. The singular value decomposition (SVD) of matrix X can be written as follows:

$$X = UDV^T,\ \ U^TU = I_p,\ \ V^TV = I_n \qquad (1)$$

The PMD generalizes this decomposition by imposing additional constraints on U and V. The rank-one PMD can be formulated as the following optimization problem:

$$\min_{d,\mathbf{u},\mathbf{v}} \frac{1}{2}\left\| X - d\mathbf{u}\mathbf{v}^T \right\|_F^2 \qquad (2)$$

$$\text{s.t. } \|\mathbf{u}\|_2^2 = 1, \|\mathbf{v}\|_2^2 = 1, P_1(\mathbf{u}) \le \alpha_1, P_2(\mathbf{v}) \le \alpha_2, d \ge 0.$$

where **u** is a column of *U*, **v** is a column of *V*, *d* is a diagonal element of *D*, $\|\bullet\|_F$ is the Frobenius norm, P_1 and P_2 are penalty functions that can take a variety of forms [24].

Let U and V be $p \times K$ and $n \times K$ orthogonal matrices, respectively, and D a diagonal matrix with diagonal elements d_k, it can be proved that [24]

$$\frac{1}{2}\left\|X - UDV^T\right\|_F^2 = \frac{1}{2}\left\|X\right\|_F^2 - \sum_{k=1}^{K}\mathbf{u}_k^T X\mathbf{v}_k d_k + \frac{1}{2}\sum_{k=1}^{K} d_k^2 \qquad (3)$$

Hence, when *K*=1, we can see that **u** and **v** satisfying Eq. (2) can also satisfying the following problem:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T X\mathbf{v} \qquad (4)$$

$$\text{s.t.} \left\|\mathbf{u}\right\|_2^2 = 1, \left\|\mathbf{v}\right\|_2^2 = 1, P_1(\mathbf{u}) \le \alpha_1, P_2(\mathbf{v}) \le \alpha_2$$

and the *d* satisfying Eq.(2) is $d = \mathbf{u}^T X\mathbf{v}$.

The optimization problem in Eq. (4) can be finessed to the following biconvex optimization [24]:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T X\mathbf{v} \qquad (5)$$

$$\text{s.t.} \left\|\mathbf{u}\right\|_2^2 \le 1, \left\|\mathbf{v}\right\|_2^2 \le 1, P_1(\mathbf{u}) \le \alpha_1, P_2(\mathbf{v}) \le \alpha_2$$

It can be turned out that the solution to Eq. (5) satisfies Eq. (4) provided that α is chosen appropriately [24].

Eq. (5) is called the rank-1 PMD, and the iterative algorithm used to optimize it is summarized as following:

Step1. Initialize **v** to have unit L_2-norm.

Step2. Iterate until convergence:

(a) $\mathbf{u} \leftarrow \arg\max_{\mathbf{u}} \mathbf{u}^T X\mathbf{v}$, s.t. $\left\|\mathbf{u}\right\|_2^2 \le 1, P_1(\mathbf{u}) \le \alpha_1$.

(b) $\mathbf{v} \leftarrow \arg\max_{\mathbf{v}} \mathbf{u}^T X\mathbf{v}$, s.t. $\left\|\mathbf{v}\right\|_2^2 \le 1,\ P_2(\mathbf{v}) \le \alpha_2$.

Step3. $d \leftarrow \mathbf{u}^T X\mathbf{v}$.

To obtain multiple factors of PMD, we can maximize the criterion in Eq. (5) repeatedly, each time using the residuals obtained by subtracting the product of previous factors $d\mathbf{uv}$ from X, i.e. $X^{k+1} \leftarrow X^k - d_k \mathbf{u}_k \mathbf{v}_k^T$. The detailed algorithm of PMD can be found in [24]. In this paper, we take the l_1-norm of **u** and **v** as the penalty function, i.e., $\|\mathbf{u}\|_1 \le \alpha_1, \|\mathbf{v}\|_1 \le \alpha_2$. By choosing appropriately the parameters α_1 and α_2, PMD can result in sparse factors **u** and **v**. Generally speaking, α_1 and α_2 should be restricted to the ranges $1 \le \alpha_1 \le \sqrt{p}$ and $1 \le \alpha_2 \le \sqrt{n}$ [24]. Examples to demonstrate the efficiency of PMD in discovering latent factors can be found in [24].

2.2 Sample Clustering using PMD

For gene expression data set, the number of genes *p* is typically in the thousands, while the number of experiments *n* is typically less than one hundred. The data are represented by an expression matrix *X* of size $p \times n$, each row of *X* containing the expression levels of a gene in all the *n* samples, while each column of *X* containing the expression levels of all the *p* genes in one sample.

Our goal is to find a small number of metasamples, each of which is defined as a linear combination of the *n* samples. On the other hand, we can approximate each sample as a linear combination of these metasamples, and consequently, we can cluster the samples according to their representations over the metasamples. In other words, the metasamples serve as the cluster centers. Mathematically, this can be accomplished by factorizing matrix *X* into two matrices: $X \sim UH$. Matrix *U* is of size $p \times k$, with each of the *k* columns defining a metasample, and each entry u_{ij} in *U* representing the expression level of gene *i* over metasample *j*. Matrix *H* is of size

$k \times n$, with each of the *n* columns representing the metasample expression pattern of the corresponding sample, and each entry h_{ij} representing the coefficient of metasample *i* over sample *j*. Figure 1 shows a simple case with $k = 2$.

Figure1 Here

We do the factorization $X \sim UDV^T = UH$ by using PMD with sparsity constraints imposed on *V*. Since *D* is a diagonal matrix, it only affects the magnitude of the non-zero elements in V^T. So matrix *H* has the same sparsity as V^T. After factorizing *X*, we can use matrix V^T to group the *n* samples into *k* clusters. Samples corresponding to the non-zero elements in each row of V^T are placed into a cluster; that is, sample *i* is placed in cluster *j* if v_{ij} is non-zero, where v_{ij} is the element of V. In the following section, we illustrate the principle of the proposed PMD based clustering.

2.3 Pattern Inference using PMD

Let $\mathbf{u}_1$ and $\mathbf{v}_1$ be the first pair of factors extracted from *X* by using PMD with $P_2(\mathbf{v}_1) = \|\mathbf{v}_1\|_1 \le \alpha_2$ but without constraint P_1 on $\mathbf{u}_1$. By choosing an appropriate α_2, we can get a sparse vector $\mathbf{v}_1$ with many entries being (nearly) zero. Without loss of generality, suppose that the first c_1 entries in $\mathbf{v}_1$ are non-zero, i.e.,

$$\mathbf{v}_1 = [v_{1,1}, v_{1,2}, \cdots v_{1,c_1}, 0, \cdots 0]^T \qquad (6)$$

Then

$$\hat{X}_1 = \mathbf{u}_1 d_1 \mathbf{v}_1^T = d_1[v_{1,1}\mathbf{u}_1; v_{1,2}\mathbf{u}_1; \cdots v_{1,c_1}\mathbf{u}_1; \mathbf{o}; \cdots \mathbf{o}] \qquad (7)$$

where $\mathbf{o}$ is a *p*-dimensional column vector with all elements being zero.

From Eq. (7) we can see that $\mathbf{u}_1$ and $\mathbf{v}_1$ can only represent the first c_1 samples in *X*. In other words, only the patterns of the first c_1 samples can be expressed as the linear combinations of metasample $\mathbf{u}_1$. If the metasample $\mathbf{u}_1$ represents the pattern of the first class, then only the first c_1 samples have similar expression pattern to metasample $\mathbf{u}_1$, and hence the first c_1 samples can be clustered into one class.

To obtain the second pair of factors, i.e., $\mathbf{u}_2$ and $\mathbf{v}_2$, we first subtract $\hat{X}_1$ from *X*:

$$X_1 = X - \hat{X}_1 = [\mathbf{e}_1;\cdots\mathbf{e}_{c_1};\mathbf{x}_{c_1+1};\cdots\mathbf{x}_n] \qquad (8)$$

where $\mathbf{e}_i = \mathbf{x}_i - d_1 v_{1,i}\mathbf{u}_1$, $i \le c_1$. Ideally, the residual $\mathbf{e}_i$ should be approximately zero. Here we assume that $\mathbf{e}_i$ fits well for this instance, and later we will discuss the situation if $\mathbf{e}_i$ is not approximately zero.

Once X_1 is obtained by Eq. (8), the second pair of factors, $\mathbf{u}_2$ and $\mathbf{v}_2$, can be extracted from it. Without loss of generality, $\mathbf{v}_2$ can be denoted as

$$\mathbf{v}_2 = [0_1,\cdots 0_{c_1}, v_{2,1}, v_{2,2},\cdots v_{2,c_2}, 0_{c_1+c_2+1},\cdots 0_n]^T \qquad (9)$$

Consequently, we have

$$\hat{X}_2 = \mathbf{u}_2 d_2 \mathbf{v}_2^T = d_2[\mathbf{o}_1;\cdots\mathbf{o}_{c_1}; v_{2,1}\mathbf{u}_2; v_{2,2}\mathbf{u}_2;\cdots v_{2,c_2}\mathbf{u}_2;\mathbf{o};\cdots\mathbf{o}] \qquad (10)$$

$$X_2 = X_1 - \hat{X}_2 = [\mathbf{e}_1;\cdots\mathbf{e}_{c_1};\mathbf{e}_{c_1+1};\cdots\mathbf{e}_{c_1+c_2};\mathbf{x}_{c_1+c_2+1};\cdots\mathbf{x}_n] \qquad (11)$$

Accordingly, the metasample $\mathbf{u}_2$ represents the pattern of the second class, and we can cluster the samples corresponding to the non-zero elements in $\mathbf{v}_2$ to another class. Repeating the above procedures, we can obtain *k* pairs of factors, i.e., $\mathbf{u}_1,\cdots,\mathbf{u}_k$ and $\mathbf{v}_1,\cdots,\mathbf{v}_k$, and assign each sample in data set *X* to a class.

Let's then discuss the situation when there are some residuals $\mathbf{e}_i$ that have non-negligible values. As an example, we assume $\mathbf{e}_{c_1}$ has a relatively large Frobenius norm, i.e., the pattern of sample $\mathbf{x}_{c_1}$ cannot be perfectly represented by using only one metasample $\mathbf{u}_1$. In other words, $\mathbf{x}_{c_1}$ may be a linear combination of $\mathbf{u}_1$ and some other metasamples. Without loss of generality, assuming that $\mathbf{x}_{c_1}$ is the linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$, we have

$$\mathbf{v}'_2 = [0_1, \cdots 0_{c_1-1}, v'_{2,1}, v'_{2,2}, \cdots v'_{2,c'_2}, 0_{c_1+c'_2+1}, \cdots 0_n]^T \qquad (12)$$

where $c'_2 = c_2 + 1$. According to the principle of clustering, the sample $\mathbf{x}_{c_1}$ should be clustered to both class 1 and class 2.

2.4 Model Selection

In the PMD model of gene expression data, a key issue is how to determine the rank *k*, i.e., the number of clusters. In fact, how to determine the number of clusters is still an open problem in gene expression data clustering. Interestingly, with PMD we can determine the number of clusters by the following method. From $d = \mathbf{u}^T X\mathbf{v}$ we can see that *d* represents the distribution of the source data's energy over each of the factors. In general, *d* will monotonically decrease with the increase of *k* (refer to Figure 3 please). Since each metasample represents the pattern of a class, *d* will not decrease much when all the meaningful metasamples were extracted. In other words, through observing how *d* changes as *k* increases, we can determine *k* as the value at which *d* falls significantly, e.g. *k*=4, in Figure 3. Since in this process we only want to impose sparsity constraint on $\mathbf{v}$ but not $\mathbf{u}$, we let $\alpha_1 = \sqrt{p}$. For α_2, since it is restricted in the range $1 \le \alpha_2 \le \sqrt{n}$, and we found experimentally that when α_2 is about $0.5\sqrt{n}$ the factor $\mathbf{u}$ is sparse, we roughly let $\alpha_2 = 0.5\sqrt{n}$.

After determining *k*, we then choose a precise value for α_2. When taking a small value, e.g., $\alpha_2 = 0.3\sqrt{n}$, only part of the samples are clustered (Tables 2, 4 and 6).

With the increase of α_2, more and more samples are clustered. When every sample is clustered to at least one class, the corresponding α_2 should be the desired value.

3. Experimental Results

In this section we evaluate the proposed method by applying it to elucidate cancer subtypes and cell differentiation. Three cancer data sets, i.e., the acute leukemia data set, the central nervous system tumor data set and the SRBCT cancer data set, are used to test our method. In addition, experiment on the lymphoid development data set is used to validate the potential capacities of our method for cell differentiation analysis. The Matlab source code of our proposed method can be downloaded at http://www4.comp.polyu.edu.hk/~cslzhang/code/MPD_PMD.rar.

3.1 Experiments on the Cancer Data Sets

3.1.1 Leukemia Data Set

This data set contains *p*=5000 genes in 38 samples, and it consists of 19 cases of B_cell acute lymphoblastic leukemia (ALL_B), 8 cases of T_cell acute lymphoblastic leukemia (ALL_T) and 11 cases of acute myelogenous leukemia (AML). In this data set, the distinction between acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL), as well as the division of ALL into T and B cell subtypes, are known.

This data set is well established and it is a benchmark data set for comparing the performance of different clustering algorithms. In general, most clustering algorithms work well on this data set. For example, HC can have good clustering results with appropriate choices of linage metric and the number of input genes [5]. However, HC is unstable because its performance is subject to the number of input genes. It can correctly find the AML–ALL distinction only when the average linkage metric is used and the number of input genes is between 1,800 and 3,200 (the only incorrect assignment involves one of the known outlier samples). However, HC cannot correctly find the important distinction between ALL-T and ALL-B.

SOM could also reveal the distinctions on this data set [18]. Golub et al. [10] found that, for 2 classes, SOM may split the data into [AML] vs. [ALL-T+ALL-B] or into [AML+ALL-T] vs. [ALL-B], depending on the initial conditions.

We also applied two recently developed popular clustering methods, the AP [31] and SC [34], to this data set. The results are listed in the Supplemental Tables S1 (for 2 classes) and S2 (for 3 classes), which were published on the IEEE web site. These two methods cannot find the AML–ALL distinction for 2 classes, either. Compared with HC and SOM, no advantage is embodied for AP and SC on the leukemia data set.

Brunet et al. [5] applied NMF to this data set. With rank *k*=2, NMF can consistently discover the ALL-AML biological distinction with high accuracy and robustness. It was also found that, a higher rank *k* can further partition the samples. The clusters show a nested structure as *k* increases from 2 to 4, and the nesting captures the known subtypes. For *k*=2, the two classes correspond to the ALL and AML samples. However, it misclassifies two ALL-B subtypes (ALL_14749_B-cell and ALL_7092_B-cell) to AML (Table 1). One possible explanation made by Brunet et al. [5] is the incorrect diagnosis of the samples. Brunet et al. [5] included them in the analysis but expected them to be outliers. For *k*=3, the partition reflects the distinction between ALL-T and ALL-B within the ALL class. Again, there are two misclassifications made (Table S2). The same ALL_14749_B-cell is once again incorrectly assigned to AML, and another ALL-B sample (ALL_21302_B-cell) is incorrectly assigned to ALL-T, showing some kind of instability. For *k*= 4, a fourth class appears which is deemed robust but its biological significance is unclear [5].

To improve the results of NMF, Gao and George [9] used sparse NMF (SNMF) to cluster this data set. Compared with NMF, when *k*=2 SNMF correctly classifies the two difficult ALL cases that are missed by NMF (Table S1). However, one AML sample (AML_13) is assigned to ALL. When *k*=3, SNMF well splits the ALL samples into two subtypes without mistake. However, it still misclassifies one AML sample to ALL. One possible explanation may be the incorrect diagnosis of this sample. Also, Gao and George suspected that there may exist more than three subclasses in the leukemia data set.

From the above published works we can see that, although these clustering methods (HC, SOM, AP, SC, NMF and SNMF) work well, the results are not consistent. By using HC, SOM, AP and SC, the distinction between AML and ALL cannot be found. Although NMF and SNMF can discover the ALL-AML distinction with high accuracy and robustness [5, 9, 28], the clustering results on three samples, i.e., ALL_14749_B-cell, ALL_7092_B-cell and AML_13, are not consistent. In addition, Brunet et al. [5] found a robust fourth class, and Gao and George [9] suspected that there may exist more than three subclasses in the leukemia data set.

Figures 2, 3 Here

We then applied the proposed PMD based method to this data set. Since we only want to impose sparsity on $\mathbf{v}$ but not $\mathbf{u}$, we let $\alpha_1 = \sqrt{p}$. For α_2, we let $\alpha_2 = 0.5\sqrt{n}$ (refer to the subsection ‘Model Selection’). The experimental results are given in Table 1 and Figures 2 and 3. From the two figures we can see that, identical to the results by NMF and SNMF, four subclasses, instead of three subclasses, should be selected for this data set because the value of *d* has a significant drop from *k*=4 to *k*=5 (Figure 3), implying that *k*=4 can characterize all the patterns in this data set.

Table 1 Here

The clustering results for 4 classes by all the seven methods are listed in Table S3. From Table S3 we can see that, the results of PMD can reasonably explain why the results of the other six methods are not consistent on some samples. For example, according to the clustering result of PMD, some ALL-B samples should be the combination of subclass 1 and subclass 4, and the other ALL-B samples belong to either subclass 1 or subclass 4. This is identical to the result that SOM split ALL-B samples into two groups [10] when *k*=4. In fact, AP, SC and NMF also split ALL-B samples into two groups when 4 classes are chosen. On the other hand, ALL-T samples are clustered to subclass 2 very consistently. According to the experimental results by PMD, some AML samples should be compound tumours.

Table 2 Here

To further investigate the effect of sparsity constraint on the experimental results, we perform the experiments with different values of α_2 (Table 2) by fixing α_1. When α_2 takes smaller values, i.e., with stronger sparsity constraint, only a small portion of the samples will be clustered. With the increase of α_2, more and more samples could be clustered to certain clusters, and the results are consistent.

From the above analysis we can see that HC, SOM, AP, SC, NMF and PMD are all useful for clustering the tumour samples. Compared with the other methods, however, PMD can locate the samples that have compound subclasses. This will be very helpful in subclass discovery.

3.1.2 Central Nervous System Tumors

This data set is composed of five types of central nervous system embryonal tumours [16]. It contains $p = 5597$ genes in 42 samples, representing five distinct morphologies: 10 classic medulloblasomas (MD), 10 malignant gliomas (MGlio), 10 rhabdoids (Rhab), 4 normal cerebella (Ncer) and 8 primitive neuroectodermal tumours (PNET).

On this data set, Brunet et al. only analyzed the first four types of tumours since they found that the 8 primitive nueroectodermal tumours did not form a distinct tight class or subclass using either supervised or unsupervised clustering method [5]. In their experiments, it was found that NMF is more suitable to cluster this data set than HC and SOM.

Table 3 Here

Figures 4, 5 Here

We then cluster the whole data set using PMD. The results are given in Table 3. Figure 4 shows the image of matrix V^T with k=5. Figure 5 shows the changes of d with respect to the rank k ($\alpha_2 = 0.45\sqrt{n}$). From Figure 5 we can see that, five subclasses should be reasonable for this data set. Table S4 lists the results by the other six methods with 5 clusters. Consistent with the conclusion of Brunet et al., the 8 PNET samples do not form a distinct tight class, but they distribute over the other four classes. The experimental results of PMD show that except for Brain_PNET_7, all the other 7 PNET samples are compound tumours, which may be a reasonable interpretation for why they can not form a distinct tight class. It can also be found that

AP and SC split the MGlio samples into 2 subclasses. In summary, the PMD can well explain the disagreement of clustering results by the other methods.

Table 4 Here

We also performed the experiments with different values of α_2 on this data set. The results are listed in Table 4. Except for a couple of samples, such as Brain_MD_12 and Brain_MD_61, the results are consistent.

3.1.3 SRBCT Cancer Data Set

This data set has 63 samples with 2308 genes and 4 expression diagnosis patterns (Kahn et al., 2001). For this data set, Leone et al. reported that the best tuning-robust estimate of AP partitions this data set into 5 clusters, while making as many as 22 errors [30].

Table 5 Here

Tables 5 and S5 show the clustering results by different methods. Figure 6 shows the image of matrix V^T with *k*=5 and Figure 7 shows the changes of *d* with respect to the rank *k* ($\alpha_2 = 0.42\sqrt{n}$). From Figure 7 we can see that, five subclasses should be reasonable for this data set, which is consistent with the result of AP [30]. The experimental results of PMD also show that many samples are compound tumors, which can well explain the disagreement of clustering results on some samples by the other methods. At the same time, the clustering results of some samples, e.g., the last one in Table 5, may not be reasonable. This implies that our method still has much room to be improved.

Figures 6, 7 Here

Table 6 Here

The experimental results of PMD with different values of α_2 are listed in Table 6. Similar to the first two experiments, except for a couple of samples, the results are consistent.

3.2 Experiments on the Lymphoid Development Data Set

In all multi-cellular organisms, somatic differentiated cells are developed from embryonic stem cells in the formation phase, and from adult tissue-specific stem cells in the adult phase. The study of triggers and molecular programs that drive cells through proliferation and differentiation stages is a key issue of developmental biology. In classical models of such processes, external or internal factors will initiate and drive differentiation stages in a non-reversible manner. Diagrams can be depicted to resemble genealogies of developmental stages, which are often called developmental trees [6]. Recently, the gene expression programs of developmental trees have been studied extensively using microarrays, which help to elucidate the underlying molecular processes [1, 4, 8, 13].

In this paper, our purpose is not to infer the developmental trees using the gene expression data of developmental stages. Since the gene expression pattern of conterminous developmental stages may be analogous, here we use PMD to discover the modules of conterminous developmental stages, which may be useful for developmental biology research potentially.

Figure 8 Here

Lymphoid development has been extensively studied. Many developmental stages are known, and there is a large amount of data available on distinct stages of development and in several cell lineages (Figure 8). In this paper, the lymphoid development data set contains four stages of early development hematopoietic cells [1] (hematopoietic stem cell (HSC), multipotent progenitor (MPP), common lymphoid progenitor (CLP) and common myeloid progenitor (CMP)); three B-cell lineage stages [21] (pro-B cells (Bpro), pre-B cells (Bpre) and immature B-cells (Bimm)); one natural killer (NK) stage [17]; and four T-cell lineage stages [26] (double negative T-cells (TDN), cd4 T cells (TCD4), cd8 T-cells (TCD8) and natural killer T-cells (TNK)). The developmental tree describing the order of differentiation of the cells is shown in Figure 8. The final data set consists of 11 developmental stages and 3697 genes (HSC was used as reference when pre-processing the data set). We got the data from [6], where the pre-processing procedures can also be found.

In this experiment, we set $\alpha_2 = 0.55\sqrt{n}$ when applying PMD to the data set. The experimental results are given in Figures 9 and 10. From Figure 10 we can see that, there should be three clusters for this data set. Figure 9 shows the samples of the three clusters, which correspond to the three groups in Figure 8. From Figures 8 and 9 we see that each cluster corresponds to a branch of the development tree. Especially, stage 2 (CLP) is clustered into two groups since it is the crotch in the development tree.

Figures 9, 10 Here

4. Conclusions and Discussions

In this study, we proposed to use the penalized matrix decomposition (PMD) to extract metasamples from gene expression data. With the sparsity constrain on the decomposition factors, the extracted metasamples can well capture the intrinsic structures of the samples in the same class. Meanwhile, the PMD factors of each sample are good indicators of the class label of it. Compared with traditional methods, such as HC, SOM, AP, SC and NMF, the proposed method can identify the samples with complex classes. The experimental results on four representative data sets showed that the proposed method is able to effectively discover biological phenotypes, verifying that PMD is a powerful tool for gene expression data clustering.

It should be mentioned that we found experimentally that *d* can be used to determine the number of clusters according to its changing tendency. When *d* falls significantly from *k* to *k*+1, this means that all the meaningful patterns can be extracted using *k* clusters. If *d* has a gradual decay from *k* to *k*+1, this implies that more than *k* meaningful patterns may exist in the dataset. This is the reason that why *d* can be used to determine the number of clusters according to its changing tendency. However, at present, we can not conclude in theory that *d* must be or must not be the indicative value. It needs more investigation in the future. Fortunately, there have been some similar works on the statistical significance of matrix eigenvalues [32,33,35], which may be useful to our future study.

Although DNA microarray technology is a potential method for disease diagnosis, especially for gene related diseases, it can be found that for some samples, different methods may cluster them into different subclasses. Therefore, currently it is more appropriate to serve as an assistant technology. Interestingly, it can be found that the proposed PMD method provides a reasonable explanation on these inconsistent classifications by various methods such as HC, SOM, AP, SC and NMF, etc. It can find more than one subclasses contained in one sample. A challenging work in the future is how to provide a meaningful biological interpretation of the classes discovered by PMD when the class labels and substructures of the data set are unknown. In addition, how to introduce the biological interpretation into the metasample calculation process is another problem that deserves further study.

Finally, it should be noted that there is a dual view of decomposition $X \sim UDV^T$, which defines metagenes (rows of *V*) and clusters the genes according to the entries of *U*. One can study the factor *U* for pathway enrichment analysis or other interpretations of biological significance. We do not focus on this view in this paper, but it is clearly of great interest. A good example of using factor models for interrogating biological pathways can be found in [39]. We will further study it in future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 30700161& 30900321 and the Foundation for Young Scientist of Shandong Province, China under Grant No. 2008BS01010.

References

1. Akashi K, He X, Chen J, Iwasaki H, Niu C, Steenhard B, Zhang J, Haug J, Li L: Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis. Blood 2003, 101:383–389.
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X. et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000, 403: 503–511.

3. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl Acad. Sci. USA 1999, 96:6745–6750.
4. Anisimov SV, et al.: ‘NeuroStem Chip’: a novel highly specialized tool to study neural differentiation pathways in human stem cells. BMC Genomics 2007, 8: 46.
5. Brunet JP, Tamayo P, Golun TR, Mesirov JP: Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 2004, 101(12): 4164–4169.
6. Costa IG, Roepcke S, Hafemeister C, Schliep, A: Inferring differentiation pathways from gene expression. Bioinformatics 2008, 24(13):i156-i164.
7. Eisen MB, Spellman PT, Brown PO, Botstein D, et al.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA 1998, 95:14863-14868.
8. Ferrari F, Bortoluzzi S, Basso D, Bicciato S, Zini R, Gemelli C, Danieli GA, Ferrari S, Genomic expression during human myelopoiesis. BMC Genomics 2007, 8:264.
9. Gao Y, George C: Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 2005, 21:3970-3975.
10. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286: 531–537.

11. Houck K, Nikrui N, Duska L, Chang Y, Fuller AF, Bell D, Goodman A, Borderline tumors of the ovary: correlation of frozen and permanent histopathologic diagnosis. Obstet Gynecol 2000, 95:839-43.

12. Huang DS, Zheng CH: Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics 2006, 22: 1855-1862.

13. Hyatt G, Melamed R, Park R, Seguritan R, Laplace C, Poirot L, Zucchelli S, Obst R.: Gene expression microarrays: glimpses of the immunological genome. Nat. Immunol. 2006, 7: 686–691.

14. Okumi M, Matsuoka Y, Tsukikawa M, Fujimoto N, Sagawa S, Itoh K: A compound tumor in the adrenal medulla-pheochromocytoma combined with ganglioneuroma: a case report. Acta Urologica Japonica 2000, 46:887-890.

15. Perou C M, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al.: Molecular portraits of human breast tumours, Nature 2000, 406: 747–752.

16. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, et al.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002, 415:436–442.

17. Poirot L, Poirot L, Benoist C, Mathis D, et al.: Natural killer cells distinguish innocuous and destructive forms of pancreatic islet autoimmunity. Proc. Natl Acad. Sci. USA 2004, 101:8102–8107.

18. Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES: Class prediction and discovery using gene expression data. In: Proceedings of the Fourth

International Conference on Computational Molecular Biology, Tokyo, Japan, RECOMB 2000, 263–272.

19. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 1998, 9:3273–3297.
20. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al.: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA 1999, 96:2907-2912.
21. Tze LE, Schram BR, Lam KP, Hogquist KA, Hippen KL, Liu J, et al.: Basal immunoglobulin signaling actively maintains developmental stage in immature B cells. PLoS Biol 2005, 3:e82.
22. Wang HQ, Wong HS, Huang DS, Shu J, Extracting gene regulation information for cancer classification. Pattern Recognition 2007, 40:3379-33927.
23. Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O: Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics 2002, 3:36.
24. Witten DM, Tibshirani R, Hastie T: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009, 10(3):515-534.
25. Witten DM, Tibshirani R: Extensions of sparse canonical correlation analysis, with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 2009, 8(1): Article 28.

26. Yamagata T, Benoist C, Mathis D: A shared gene-expression signature in innate-like lymphocytes. Immunol. Rev. 2006, 210:52–66.

27. Wang H, Zheng H, Azuaje F: Poisson-based self-organizing feature maps and hierarchical clustering for serial analysis of gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2007, 4(2):163-175.

28. Zheng CH, Huang DS, Zhang L, Kong XZ: Tumor clustering using non-negative matrix factorization with gene selection. IEEE Transactions on Information Technology in Biomedicine 2009, 13(4):599-607.

29. Paul TK, Iba H: Prediction of cancer class with majority voting genetic programming classifier using gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2009, 6(2):353 – 367.

30. Leone M, Sumedha, Weight M: Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics 2007, 23:2708-2715.

31. Frey JF, Dueck D. Clustering by passing messages between data points. Science 2007, 315:972–976.

32. Bouchaud JP, Potters M: Financial applications of random matrix theory: a short review. http://arxiv.org/abs/0910.1205

33. Karoui NE: Spectrum estimation for large dimensional covariance matrices using random matrix theory. http://www.stat.berkeley.edu/~nkaroui/papers/AOS581SpectrumEstimationRMT.pdf

34. Ng AY, Jorden MI, Weiss Y: On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems 2002, 14:849-856.

35. Halabi N, Rivoire O, Leibler S, Ranganathan R: Protein sectors: evolutionary units of three-dimensional structure. Cell 2009, 138(4): 774-786.

36. Zhao XM, Cheung YM, Huang DS: Analysis of Gene Expression Data Using RPEM Algorithm in Normal Mixture Model with Dynamic Adjustment of Learning Rate. International Journal of Pattern Recognition and Artificial Intelligence 2010, 24(4): 651-666.

37. Li H, Sun Y, Zhan M: The discovery of transcriptional modules by a two-stage matrix decomposition approach, Bioinformatics 2007 23(4):473-479.

38. Zhao XM, Wang RS, Chen LN, and Aihara Kazuyuki: Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucl. Acids Res 2008, 36(9):e48.

39. Chang JT, Carvalho C, Mori S, Bild AH, Gatza ML, Wang Q, Lucas JE, Potti A, Febbo PG, West M, Nevins JR: A genomic strategy to elucidate modules of oncogenic pathway signaling networks. Mol Cell 2009, 34(1): 104-14.

Figures

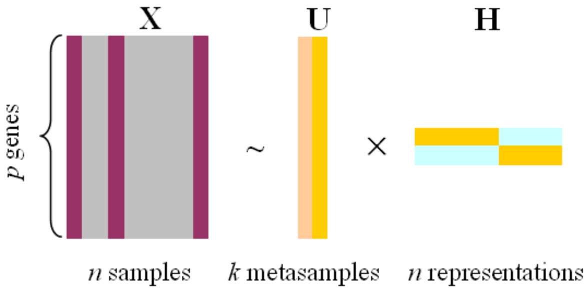

Figure 1. A rank-2 reduction of gene expression data using matrix factorization.

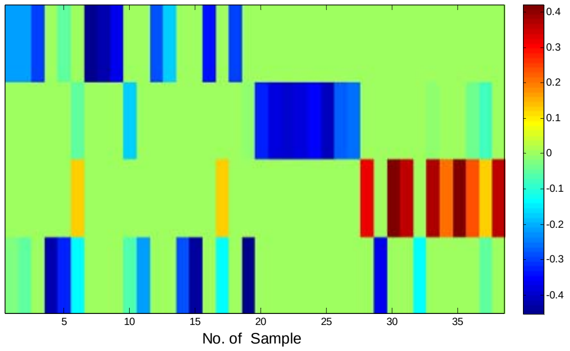

Figure 2. The image of matrix V^T extracted from the leukemia data set, which is used to cluster the samples. The 1st row of V^T is the first factor $\mathbf{v}_1$, the 2nd row of V^T is $\mathbf{v}_2$, and so on.

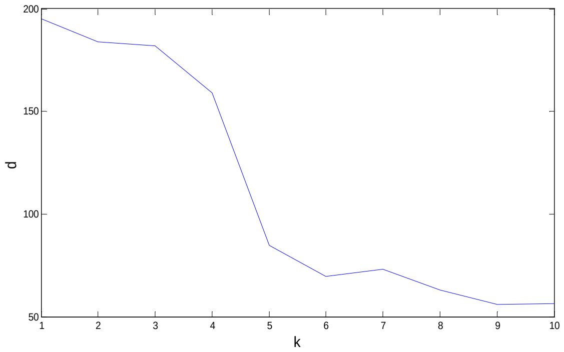

Figure 3. The value of *d* with respect to *k*. When *k* is changed from 4 to 5, there is a significant drop of the value of *d*. This suggests that 4 should be the number of clusters.

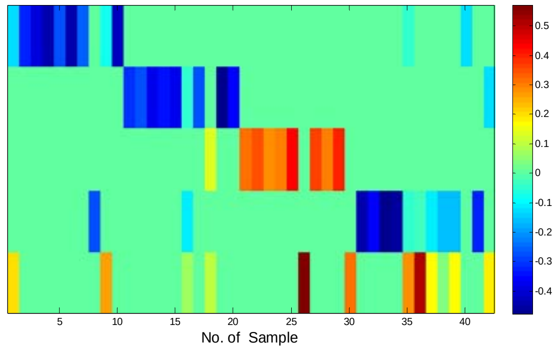

Figure 4. The image of matrix V^T extracted from the central nervous system embryonal tumours data set, which is used to cluster the samples. The 1st row of V^T is the first factor $\mathbf{v}_1$, the 2nd row is $\mathbf{v}_2$, and so on.

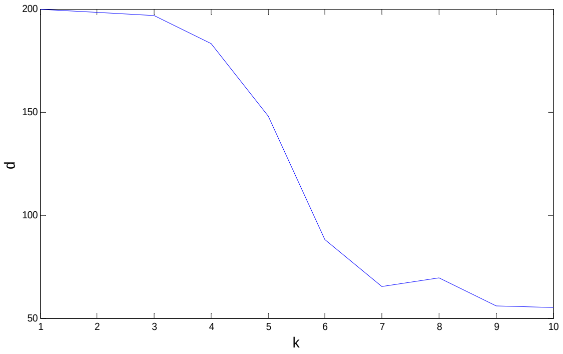

Figure 5. The value of *d* with respect to *k*. When *k* is changed from 5 to 6, there is a significant drop of the value of *d*. This suggests that 5 should be the number of clusters.

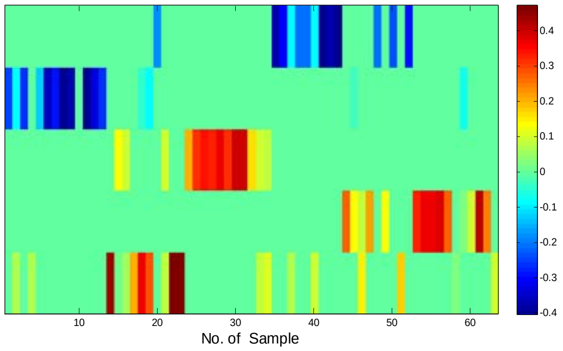

Figure 6. The image of matrix V^T extracted from the lymphoid development data set, which is used to cluster the samples. The 1st row of V^T is the first factor $\mathbf{v}_1$, the 2nd row is the $\mathbf{v}_2$, and so on.

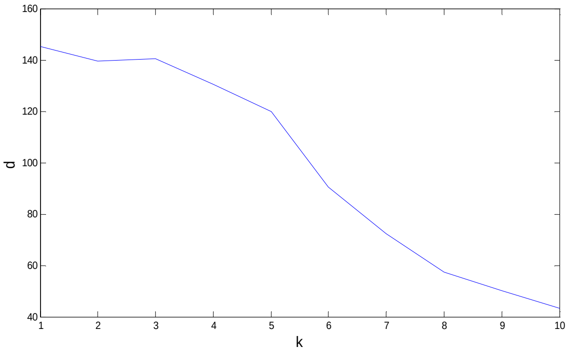

Figure 7. The value of *d* with respect to *k*. When *k* is changed from 5 to 6, there is a significant drop of the value of *d*. This suggests that 5 should be the number of clusters.

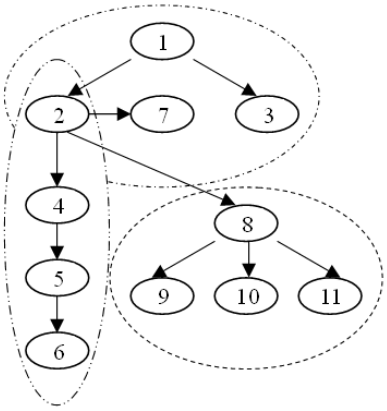

Figure 8. The developmental tree with the stages contained in the Lymphoid data set. Three groups in dash ellipse are the clusters found by using PMD. The numbers of the developmental stages are as follows: 1for MPP, 2 for CLP, 3 for CMP, 4 for Bpro, 5 for Bpre, 6 for Bimm, 7 for NK, 8 for TDN, 9 for TCD4, 10 for TCD8, and 11 for TNK.

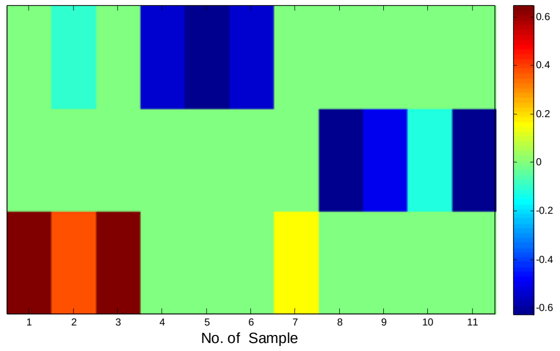

Figure 9. The image of matrix V^T extracted from the lymphoid development data set, which is used to cluster the samples. The 1st row of V^T is the first factor $\mathbf{v}_1$, the 2nd row is the $\mathbf{v}_2$, and so on.

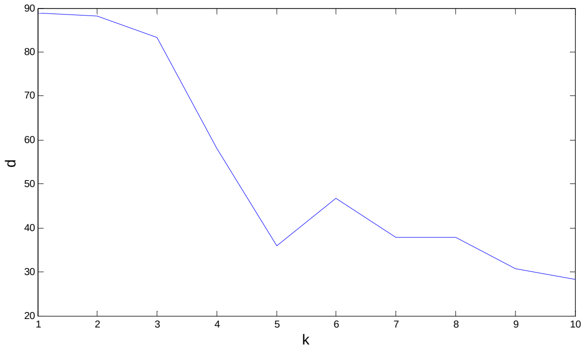

Figure 10. The value of *d* with respect to *k*. When *k* is changed from 3 to 4, there is a significant drop of the value of *d*. This suggests that 3 should be the number of clusters.

Tables

Table 1. Class assignment of the leukemia data set based on PMD(4 clusters).

No.	Sample Label	Cluster	No.	Sample Label	Cluster
1	ALL_B	1,4	20	ALL_T	2
2	ALL_B	1,4	21	ALL_T	2
3	ALL_B	1,4	22	ALL_T	2
4	ALL_B	4	23	ALL_T	2
5	ALL_B	1,4	24	ALL_T	2
6	ALL_B	2,3,4	25	ALL_T	2
7	ALL_B	1	26	ALL_T	2
8	ALL_B	1	27	ALL_T	2
9	ALL_B	1	28	AML	3
10	ALL_B	2,4	29	AML	4
11	ALL_B	4	30	AML	3
12	ALL_B	1	31	AML	3
13	ALL_B	1	32	AML	4
14	ALL_B	4	33	AML	2,3
15	ALL_B	4	34	AML	3
16	ALL_B	1	35	AML	3
17	ALL_B	3,4	36	AML	2,3
18	ALL_B	1	37	AML	2,3,4
19	ALL_B	2,4	38	AML	3

Table 2 Class assignment using PMD with different α_2 (4 clusters).

No.	Sample label	$\alpha_2\,(\times\sqrt{n}\,)$					No.	Sample label	$\alpha_2\,(\times\sqrt{n}\,)$				
		0.3	0.35	0.4	0.45	0.5			0.3	0.35	0.4	0.45	0.5
1	ALL_B	1	1	1	1,4	1,4	20	ALL_T			2	2	2
2	ALL_B	4	4	1,4	1,4	1,4	21	ALL_T		2	2	2	2
3	ALL_B				1	1,4	22	ALL_T	2	2	2	2	2
4	ALL_B	4	4	4	4	4	23	ALL_T	2	2	2	2	2
5	ALL_B			4	4	1,4	24	ALL_T	2	2	2	2	2
6	ALL_B		4	3,4	3	2,3,4	25	ALL_T	2	2	2	2	2
7	ALL_B	1	1	1	1	1	26	ALL_T				2	2
8	ALL_B		1	1	1	1	27	ALL_T	2		2	2	2
9	ALL_B	1	1	1	1	1	28	AML			3	3	3
10	ALL_B				4	2,4	29	AML		4	4	4	4
11	ALL_B					4	30	AML	3	3	3	3	3
12	ALL_B	1	1	1	1	1	31	AML	3	3	3	3	3
13	ALL_B				1	1	32	AML					4
14	ALL_B				4	4	33	AML		3	3	3	2,3
15	ALL_B	4	4	4	4	4	34	AML					3
16	ALL_B			1	1	1	35	AML	3	3	3	3	3
17	ALL_B				4	3,4	36	AML			3	3	2,3
18	ALL_B	1	1	1	1	1	37	AML		3	3	3	2,3,4
19	ALL_B	4	4	4	4	2,4	38	AML	3	3	3	3	3

Table 3. Class assignment of the central nervous system embryonal tumor data set (5 clusters).

No.	Sample Label	Cluster	No.	Sample Label	Cluster
1	Brain_MD	1,5	22	Brain_Rhab	3
2	Brain_MD	1	23	Brain_Rhab	3,4
3	Brain_MD	1	24	Brain_Rhab	3
4	Brain_MD	1	25	Brain_Rhab	3
5	Brain_MD	1	26	Brain_Rhab	5
6	Brain_MD	1	27	Brain_Rhab	3
7	Brain_MD	1	28	Brain_Rhab	3
8	Brain_MD	4	29	Brain_Rhab	3
9	Brain_MD	1,5	30	Brain_Rhab	5
10	Brain_MD	1	31	Brain_Ncer	4
11	Brain_MGlio	2	32	Brain_Ncer	4
12	Brain_MGlio	2	33	Brain_Ncer	4
13	Brain_MGlio	2	34	Brain_Ncer	4
14	Brain_MGlio	2	35	Brain_PNET	1,4,5
15	Brain_MGlio	2	36	Brain_PNET	3,4,5
16	Brain_MGlio	2,4,5	37	Brain_PNET	4,5
17	Brain_MGlio	2	38	Brain_PNET	4,5
18	Brain_MGlio	3,5	39	Brain_PNET	4,5
19	Brain_MGlio	2	40	Brain_PNET	1,4
20	Brain_MGlio	2	41	Brain_PNET	4
21	Brain_Rhab	3	42	Brain_PNET	2,5

Table 4. Class assignment using PMD with different α_2 (5 clusters).

No.	Sample label	$\alpha_2\,(\times\sqrt{n})$				No.	Sample label	$\alpha_2\,(\times\sqrt{n})$			
		0.3	0.35	0.4	0.45			0.3	0.35	0.4	0.45
1	Brain_MD	3		1,3	1,5	22	Brain_Rhab		3	3	3
2	Brain_MD	5	1	1	1	23	Brain_Rhab		3,5	3	3,4
3	Brain_MD	1	1	1	1	24	Brain_Rhab		5	3	3
4	Brain_MD	1	1	1	1	25	Brain_Rhab	3	3	3	3
5	Brain_MD	1,5	1	1	1	26	Brain_Rhab			5	5
6	Brain_MD	1	1	1	1	27	Brain_Rhab	3	3	3	3
7	Brain_MD	5	5	1	1	28	Brain_Rhab	3	3	3	3
8	Brain_MD		4	4	4	29	Brain_Rhab	3	3	3	3
9	Brain_MD		5	1,5	1,5	30	Brain_Rhab				5
10	Brain_MD	1	1	1	1	31	Brain_Ncer	4	4	4	4
11	Brain_MGlio	2	2	2	2	32	Brain_Ncer	4	4	4	4
12	Brain_MGlio	2		2	2	33	Brain_Ncer	4	4	4	4
13	Brain_MGlio	2	2	2	2	34	Brain_Ncer	4	4	4	4
14	Brain_MGlio	2	2	2	2	35	Brain_PNET	5	5	5	1,4,5
15	Brain_MGlio		2	2	2	36	Brain_PNET		5	5	3,4,5
16	Brain_MGlio			4,5	2,4,5	37	Brain_PNET			5	4,5
17	Brain_MGlio		2,5	2	2	38	Brain_PNET			5	4,5
18	Brain_MGlio		5	5	3,5	39	Brain_PNET	5	5	5	4,5
19	Brain_MGlio	2	2	2	2	40	Brain_PNET	5	4	4	1,4
20	Brain_MGlio	2	2	2	2	41	Brain_PNET		4	4	4
21	Brain_Rhab	3	3	3	3	42	Brain_PNET		5	4,5	2,5

Table 5. Class assignment of the SRBCT data set (5 clusters).

No.	Sample label	Cluster	No.	Sample label	Cluster	No.	Sample label	Cluster
1	EWS_T	2	22	EWS_C	5	43	NB_C	1
2	EWS_T	2,5	23	EWS_C	5	44	RMS_C	4
3	EWS_T	2	24	BL_C	3	45	RMS_C	2,4
4	EWS_T	5	25	BL_C	3	46	RMS_C	4
5	EWS_T	2	26	BL_C	3	47	RMS_C	2,4
6	EWS_T	2	27	BL_C	3	48	RMS_C	1
7	EWS_T	2	28	BL_C	3	49	RMS_C	4
8	EWS_T	2	29	BL_C	3	50	RMS_C	1
9	EWS_T	2	30	BL_C	3	51	RMS_C	5
10	EWS_T	5	31	BL_C	3	52	RMS_C	1
11	EWS_T	2	32	NB_C	3	53	RMS_C	4
12	EWS_T	2	33	NB_C	3,5	54	RMS_T	4
13	EWS_T	2	34	NB_C	3,5	55	RMS_T	4
14	EWS_C	5	35	NB_C	1	56	RMS_T	4
15	EWS_C	3,5	36	NB_C	1	57	RMS_T	4
16	EWS_C	3,5	37	NB_C	1,5	58	RMS_T	4,5
17	EWS_C	5	38	NB_C	1	59	RMS_T	2,4
18	EWS_C	2,5	39	NB_C	1	60	RMS_T	4
19	EWS_C	2,5	40	NB_C	1,5	61	RMS_T	4
20	EWS_C	1	41	NB_C	1	62	RMS_T	4
21	EWS_C	3,5	42	NB_C	1	63	RMS_T	5

Table 6. Class assignment using PMD with different α_2 (5 clusters).

No.	Sample label	$\alpha_2(\times\sqrt{n})$				No.	Sample label	$\alpha_2(\times\sqrt{n})$			
		0.3	0.35	0.40	0.42			0.3	0.35	0.4	0.42
1	EWS_T		2	2	2	33	NB_C		5	3,5	3,5
2	EWS_T			2	2,5	34	NB_C		5	3,5	3,5
3	EWS_T		2	2	2	35	NB_C	1	1	1	1
4	EWS_T				5	36	NB_C	1	1	1	1
5	EWS_T			2	2	37	NB_C		5	1,5	1,5
6	EWS_T	2	2	2	2	38	NB_C	1	1	1	1
7	EWS_T	2	2	2	2	39	NB_C		1	1	1
8	EWS_T	2	2	2	2	40	NB_C		5	1,5	1,5
9	EWS_T	2	2	2	2	41	NB_C	1	1	1	1
10	EWS_T				5	42	NB_C	1	1	1	1
11	EWS_T	2	2	2	2	43	NB_C	1	1	1	1
12	EWS_T	2	2	2	2	44	RMS_C	4	4	4	4
13	EWS_T	2	2	2	2	45	RMS_C	4	4	4	2,4
14	EWS_C	5	5	5	5	46	RMS_C		4,5	5	4
15	EWS_C		5	3	3,5	47	RMS_C	4	4	4	2,4
16	EWS_C			3	3,5	48	RMS_C	4	1	1	1
17	EWS_C			5	5	49	RMS_C		5	4,5	4
18	EWS_C	5	5	2,5	2,5	50	RMS_C	4	1	1	1
19	EWS_C	5	5	2,5	2,5	51	RMS_C			5	5
20	EWS_C	5	1,5	1	1	52	RMS_C	1	1	1	1
21	EWS_C		5	3,5	3,5	53	RMS_C		4	4	4
22	EWS_C	5	5	5	5	54	RMS_T			4	4
23	EWS_C	5	5	5	5	55	RMS_T	4	4	4	4
24	BL_C		3	3	3	56	RMS_T			4	4
25	BL_C	3	3	3	3	57	RMS_T	4	4	4	4
26	BL_C	3	3	3	3	58	RMS_T				4,5
27	BL_C	3	3	3	3	59	RMS_T	4	4	2,4	2,4
28	BL_C	3	3	3	3	60	RMS_T			5	4
29	BL_C	3	3	3	3	61	RMS_T		4	4	4
30	BL_C	3	3	3	3	62	RMS_T	4	4	4	4
31	BL_C	3	3	3	3	63	RMS_T				5
32	NB_C		5	3	3						

Supplemental Tables

Table S1. Class assignment of the leukemia dataset(2 clusters).

No.	Name of Samples	HC	SOM*	AP	SC	NMF	SNMF	PMD
1	ALL_19769_B-cell	1	1	1	1	1	1	1,4
2	ALL_23953_B-cell	1	1	1	1	1	1	1,4
3	ALL_28373_B-cell	1	1	1	1	1	1	1,4
4	ALL_9335_B-cell	1	1	1	1	1	1	4
5	ALL_9692_B-cell	1	1	1	1	1	1	1,4
6	ALL_14749_B-cell	1	2	1	1	2	1	2,3,4
7	ALL_17281_B-cell	1	1	1	1	1	1	1
8	ALL_19183_B-cell	1	1	1	1	1	1	1
9	ALL_20414_B-cell	1	1	1	1	1	1	1
10	ALL_21302_B-cell	2	1	1	1	1	1	2,4
11	ALL_549_B-cell	1	2	1	1	1	1	4
12	ALL_17929_B-cell	1	1	1	1	1	1	1
13	ALL_20185_B-cell	1	1	1	1	1	1	1
14	ALL_11103_B-cell	1	1	1	1	1	1	4
15	ALL_18239_B-cell	1	1	1	1	1	1	4
16	ALL_5982_B-cell	1	1	1	1	1	1	1
17	ALL_7092_B-cell	1	2	1	1	2	1	3,4
18	ALL_R11_B-cell	1	1	1	1	1	1	1
19	ALL_R23_B-cell	1	1	1	1	1	1	2,4
20	ALL_16415_T-cell	2	2	2	2	1	1	2
21	ALL_19881_T-cell	2	2	2	2	1	1	2
22	ALL_9186_T-cell	2	2	2	2	1	1	2
23	ALL_9723_T-cell	2	2	2	2	1	1	2
24	ALL_17269_T-cell	2	2	2	2	1	1	2
25	ALL_14402_T-cell	2	2	2	2	1	1	2
26	ALL_17638_T-cell	2	2	2	2	1	1	2
27	ALL_22474_T-cell	2	2	2	2	1	1	2
28	AML_12	1	1	1	1	2	2	3
29	AML_13	1	1	1	1	2	1	4
30	AML_14	1	1	1	1	2	2	3
31	AML_16	1	1	1	1	2	2	3
32	AML_20	1	1	1	1	2	2	4
33	AML_1	1	1	1	1	2	2	2,3
34	AML_2	1	1	1	1	2	2	3
35	AML_3	1	1	1	1	2	2	3
36	AML_5	1	1	1	1	2	2	2,3
37	AML_6	1	1	1	1	2	2	2,3,4
38	AML_7	1	1	1	1	2	2	3

*The result of SOM is not stable. It depends on the initial conditions.

Table S2. Class assignment of the leukemia dataset (3 clusters).

No.	Name of Samples	HC	SOM	AP	SC	NMF	SNMF	PMD
1	ALL_19769_B-cell	1	1	1	1	1	1	1,4
2	ALL_23953_B-cell	1	1	1	1	1	1	1,4
3	ALL_28373_B-cell	1	1	1	1	1	1	1,4
4	ALL_9335_B-cell	1	1	1	1	1	1	4
5	ALL_9692_B-cell	1	1	1	1	1	1	1,4
6	ALL_14749_B-cell	3	3	3	3	3	1	2,3,4
7	ALL_17281_B-cell	1	1	1	1	1	1	1
8	ALL_19183_B-cell	1	1	1	1	1	1	1
9	ALL_20414_B-cell	1	1	1	1	1	1	1
10	ALL_21302_B-cell	2	2	1	1	2	1	2,4
11	ALL_549_B-cell	1	1	1	1	1	1	4
12	ALL_17929_B-cell	1	1	1	1	1	1	1
13	ALL_20185_B-cell	1	1	1	1	1	1	1
14	ALL_11103_B-cell	1	1	1	1	1	1	4
15	ALL_18239_B-cell	1	1	1	1	1	1	4
16	ALL_5982_B-cell	1	1	1	1	1	1	1
17	ALL_7092_B-cell	3	3	3	3	1	1	3,4
18	ALL_R11_B-cell	1	1	1	1	1	1	1
19	ALL_R23_B-cell	1	1	1	1	1	1	2,4
20	ALL_16415_T-cell	3	2	3	2	2	2	2
21	ALL_19881_T-cell	2	2	2	2	2	2	2
22	ALL_9186_T-cell	2	2	2	2	2	2	2
23	ALL_9723_T-cell	2	2	2	2	2	2	2
24	ALL_17269_T-cell	2	2	2	2	2	2	2
25	ALL_14402_T-cell	2	2	2	2	2	2	2
26	ALL_17638_T-cell	2	2	2	2	2	2	2
27	ALL_22474_T-cell	2	2	2	2	2	2	2
28	AML_12	3	3	3	3	3	3	3
29	AML_13	1	2	3	1	3	1	4
30	AML_14	3	3	3	3	3	3	3
31	AML_16	3	3	3	3	3	3	3
32	AML_20	1	3	3	3	3	3	4
33	AML_1	3	3	3	3	3	3	2,3
34	AML_2	3	3	3	3	3	3	3
35	AML_3	3	3	3	3	3	3	3
36	AML_5	3	3	3	3	3	3	2,3
37	AML_6	1	3	3	3	3	3	2,3,4
38	AML_7	3	3	3	3	3	3	3

Table S3. Class assignment of the leukemia dataset (4 clusters).

No.	Name of Samples	HC	SOM	AP	SC	NMF	SNMF	PMD
1	ALL_19769_B-cell	1	1	1	4	1	1	1,4
2	ALL_23953_B-cell	1	1	4	4	4	1	1,4
3	ALL_28373_B-cell	1	1	4	1	1	1	1,4
4	ALL_9335_B-cell	1	4	4	4	4	4	4
5	ALL_9692_B-cell	1	1	1	4	4	4	1,4
6	ALL_14749_B-cell	1	4	3	4	4	4	2,3,4
7	ALL_17281_B-cell	1	1	1	1	1	1	1
8	ALL_19183_B-cell	1	1	1	1	1	1	1
9	ALL_20414_B-cell	1	1	1	1	1	1	1
10	ALL_21302_B-cell	2	4	4	1	2	2	2,4
11	ALL_549_B-cell	4	1	1	1	1	1	4
12	ALL_17929_B-cell	1	1	1	1	1	1	1
13	ALL_20185_B-cell	1	1	1	1	1	1	1
14	ALL_11103_B-cell	1	4	1	1	1	1	4
15	ALL_18239_B-cell	1	4	4	4	4	4	4
16	ALL_5982_B-cell	1	1	4	1	1	1	1
17	ALL_7092_B-cell	1	4	3	3	1	1	3,4
18	ALL_R11_B-cell	1	4	1	4	1	1	1
19	ALL_R23_B-cell	1	4	4	4	4	4	2,4
20	ALL_16415_T-cell	2	2	3	2	2	2	2
21	ALL_19881_T-cell	2	2	2	2	2	2	2
22	ALL_9186_T-cell	2	2	2	2	2	2	2
23	ALL_9723_T-cell	2	2	2	2	2	2	2
24	ALL_17269_T-cell	2	2	2	2	2	2	2
25	ALL_14402_T-cell	2	2	2	2	2	2	2
26	ALL_17638_T-cell	2	2	2	2	2	2	2
27	ALL_22474_T-cell	2	2	2	2	2	2	2
28	AML_12	3	3	3	3	3	3	3
29	AML_13	1	4	4	4	4	4	4
30	AML_14	3	3	3	3	3	3	3
31	AML_16	3	3	3	3	3	3	3
32	AML_20	3	3	3	3	3	3	4
33	AML_1	1	4	3	3	3	3	2,3
34	AML_2	3	3	3	3	3	3	3
35	AML_3	3	3	3	3	3	3	3
36	AML_5	3	4	4	3	3	3	2,3
37	AML_6	1	4	3	3	3	3	2,3,4
38	AML_7	3	3	3	3	3	3	3

Table S4. Class assignment of the central nervous system embryonal tumor dataset (5 clusters).

No.	Name of Samples	HC	SOM	AP	SC	NMF	SNMF	PMD
1	Brain_MD_12	1	5	3	1	1	1	1,5
2	Brain_MD_61	1	1	1	1	1	1	1
3	Brain_MD_15	1	1	1	1	1	1	1
4	Brain_MD_57	1	1	1	1	1	1	1
5	Brain_MD_33	1	1	1	1	1	1	1
6	Brain_MD_64	1	1	1	1	1	1	1
7	Brain_MD_17	1	1	1	1	1	1	1
8	Brain_MD_62	1	1	4	1	1	1	4
9	Brain_MD_63	1	5	1	1	1	1	1,5
10	Brain_MD_32	1	1	1	1	1	1	1
11	Brain_MGlio_1	2	2	2	2	2	2	2
12	Brain_MGlio_2	2	2	2	2	2	2	2
13	Brain_MGlio_3	2	2	2	2	2	2	2
14	Brain_MGlio_4	2	2	2	2	2	2	2
15	Brain_MGlio_5	2	2	**5**	**5**	2	2	2
16	Brain_MGlio_6	2	2	**5**	**5**	2	2	2,4,5
17	Brain_MGlio_7	2	2	**5**	**5**	2	2	2
18	Brain_MGlio_8	3	3	**5**	**5**	5	3	3,5
19	Brain_MGlio_9	2	2	2	2	2	2	2
20	Brain_MGlio_10	2	2	2	2	2	2	2
21	Brain_Rhab_1	3	3	3	3	3	3	3
22	Brain_Rhab_2	3	3	3	3	3	3	3
23	Brain_Rhab_3	3	3	3	3	3	3	3,4
24	Brain_Rhab_4	3	3	3	3	3	3	3
25	Brain_Rhab_5	3	3	3	3	3	3	3
26	Brain_Rhab_6	3	3	3	3	5	5	5
27	Brain_Rhab_7	3	3	3	3	3	3	3
28	Brain_Rhab_8	3	3	3	3	3	3	3
29	Brain_Rhab_9	3	3	3	3	3	3	3
30	Brain_Rhab_10	5	5	3	3	5	5	5
31	Brain_Ncer_1	4	4	4	4	4	4	4
32	Brain_Ncer_2	4	4	4	4	4	4	4
33	Brain_Ncer_3	4	4	4	4	4	4	4
34	Brain_Ncer_4	4	4	4	4	4	4	4
35	Brain_PNET_1	1	5	1	1	1	1	1,4,5
36	Brain_PNET_2	3	3	3	5	5	5	3,4,5
37	Brain_PNET_3	2	2	5	5	2	5	4,5
38	Brain_PNET_4	2	2	5	5	2	3	4,5
39	Brain_PNET_5	1	5	1	1	1	5	4,5
40	Brain_PNET_6	1	1	1	1	1	1	1,4
41	Brain_PNET_7	4	4	4	4	4	4	4
42	Brain_PNET_8	2	2	5	5	2	2	2,5

Table S5. Class assignment of the SRBCT dataset (5 clusters).

No.	Name of Samples	HC	SOM	AP	SC	NMF	SNMF	PMD
1	EWS-T1	2	2	2	4	2	5	2
2	EWS-T2	2	2	3	4	5	5	2,5
3	EWS-T3	2	2	2	2	2	2	2
4	EWS-T4	5	3	2	4	5	5	5
5	EWS-T6	2	2	2	2	2	2	2
6	EWS-T7	2	2	2	2	2	2	2
7	EWS-T9	2	2	2	2	2	2	2
8	EWS-T11	2	2	2	2	2	2	2
9	EWS-T12	2	2	2	2	2	2	2
10	EWS-T13	4	4	4	2	4	2	5
11	EWS-T14	2	2	2	2	2	2	2
12	EWS-T15	2	2	2	2	2	2	2
13	EWS-T19	2	2	2	2	2	2	2
14	EWS-C1	1	5	5	5	1	1	5
15	EWS-C2	3	2	3	3	5	5	3,5
16	EWS-C3	3	2	3	3	5	5	3,5
17	EWS-C4	3	2	3	4	5	5	5
18	EWS-C6	1	5	5	5	1	1	2,5
19	EWS-C7	1	5	5	5	1	1	2,5
20	EWS-C8	1	5	5	5	1	1	1
21	EWS-C9	3	2	3	4	5	5	3,5
22	EWS-C10	1	5	5	5	1	1	5
23	EWS-C11	1	5	5	5	1	1	5
24	BL-C1	1	1	3	3	3	3	3
25	BL-C2	1	1	3	3	3	3	3
26	BL-C3	1	1	3	3	3	3	3
27	BL-C4	1	1	3	3	3	3	3
28	BL-C5	1	1	3	3	3	3	3
29	BL-C6	1	1	3	3	3	3	3
30	BL-C7	1	1	3	3	3	3	3
31	BL-C8	1	1	3	3	3	3	3
32	NB-C1	1	1	1	3	5	5	3
33	NB-C2	1	1	1	1	5	5	3,5
34	NB-C3	1	1	1	1	5	5	3,5
35	NB-C4	1	1	1	1	1	1	1
36	NB-C5	1	1	1	1	1	1	1
37	NB-C6	1	1	1	1	1	1	1,5
38	NB-C7	1	1	1	1	1	1	1
39	NB-C8	1	1	1	1	1	1	1
40	NB-C9	1	1	1	1	1	1	1,5
41	NB-C10	1	1	1	1	1	1	1
42	NB-C11	1	1	1	1	1	1	1
43	NB-C12	1	1	1	1	1	1	1
44	RMS-C2	1	5	4	1	4	4	4
45	RMS-C3	1	5	1	1	4	4	2,4
46	RMS-C4	1	5	1	1	1	1	4
47	RMS-C5	1	5	1	1	1	1	2,4
48	RMS-C6	1	1	1	1	1	1	1
49	RMS-C7	1	1	1	1	1	1	4

50	RMS-C8	1	1	1	1	1	1	1
51	RMS-C9	3	3	3	4	5	5	5
52	RMS-C10	1	1	1	1	1	1	1
53	RMS-C11	3	3	4	4	5	5	4
54	RMS-T1	3	3	3	4	5	5	4
55	RMS-T2	4	4	4	4	4	4	4
56	RMS-T3	3	3	3	4	5	5	4
57	RMS-T4	4	4	4	4	4	4	4
58	RMS-T5	4	4	4	2	4	4	4,5
59	RMS-T6	4	4	4	4	4	4	2,4
60	RMS-T7	4	4	4	4	4	4	4
61	RMS-T8	3	3	4	4	5	5	4
62	RMS-T10	4	4	4	4	4	4	4
63	RMS-T11	4	4	4	4	4	4	5