Sharma et al. BMC Bioinformatics 2017, 18(Suppl 16):547
DOI 10.1186/s12859-017-1970-8

BMC Bioinformatics

RESEARCH Open Access

2D–EM clustering approach for high-dimensional data through folding feature vectors

Alok Sharma[1,2,3,5], Piotr J. Kamola[1,2] and Tatsuhiko Tsunoda[1,2,4*]

From 16th International Conference on Bioinformatics (InCoB 2017)
Shenzhen, China. 20-22 September 2017

Abstract

Background: Clustering methods are becoming widely utilized in biomedical research where the volume and complexity of data is rapidly increasing. Unsupervised clustering of patient information can reveal distinct phenotype groups with different underlying mechanism, risk prognosis and treatment response. However, biological datasets are usually characterized by a combination of low sample number and very high dimensionality, something that is not adequately addressed by current algorithms. While the performance of the methods is satisfactory for low dimensional data, increasing number of features results in either deterioration of accuracy or inability to cluster. To tackle these challenges, new methodologies designed specifically for such data are needed.

Results: We present 2D–EM, a clustering algorithm approach designed for small sample size and high-dimensional datasets. To employ information corresponding to data distribution and facilitate visualization, the sample is folded into its two-dimension (2D) matrix form (or feature matrix). The maximum likelihood estimate is then estimated using a modified expectation-maximization (EM) algorithm. The 2D–EM methodology was benchmarked against several existing clustering methods using 6 medically-relevant transcriptome datasets. The percentage improvement of Rand score and adjusted Rand index compared to the best performing alternative method is up to 21.9% and 155.6%, respectively. To present the general utility of the 2D–EM method we also employed 2 methylome datasets, again showing superior performance relative to established methods.

Conclusions: The 2D–EM algorithm was able to reproduce the groups in transcriptome and methylome data with high accuracy. This build confidence in the methods ability to uncover novel disease subtypes in new datasets. The design of 2D–EM algorithm enables it to handle a diverse set of challenging biomedical dataset and cluster with higher accuracy than established methods. MATLAB implementation of the tool can be freely accessed online (http://www.riken.jp/en/research/labs/ims/med_sci_math or http://www.alok-ai-lab.com/).

Keywords: EM algorithm, Feature matrix, Small sample size, Transcriptome, Methylome, Cancer, Phenotype clustering

* Correspondence: tatsuhiko.tsunoda@riken.jp
[1]Center for Integrative Medical Sciences, RIKEN Yokohama, Yokohama 230-0045, Japan
[2]CREST, JST, Yokohama 230-0045, Japan
Full list of author information is available at the end of the article

Background

The cost of molecular profiling and recruiting large cohort of patients is often a prohibitive factor which results in many biomedical datasets having much higher number of features (or dimensions) d larger than sample number n (i.e., $d >> n$). This leads to a problem usually referred to as the small sample size (SSS) problem, and make it challenging to employ many state-of-the-art clustering algorithms to group the samples appropriately. Many clustering methods are based on maximum-likelihood approach or employ covariance information [1, 2]. However, when SSS problem exists, the covariance of samples becomes singular (or ill posed) and it is difficult to effectively utilize it in the application of clustering algorithms. This restricts us to the approaches which mainly employ norm distance (e.g. Euclidean norm) or centroid of samples to categorize samples into various clusters. Examples for such kind of algorithms are k-means or hierarchical clustering (which employs norm distance to build a dendrogram) [2].

In the literature, k-means clustering algorithm received widespread attention and has been used in a range of biological applications. The underlying functionality of many of the recent tools used in multiomics data analysis (iCluster, and iClusterPlus [3]) or cancer analysis (ConsensusCluster (CC) and CCPlus [4, 5]) was built using k-means. Though this type of method has been widely applied in the literature due to its easiness and appropriate level of clustering accuracy, it does not cluster based on data distribution as covariance information has not been employed. If we can gather more information from a limited amount of data then the clustering performance can be improved. This would have consequences in findings of biological sciences, especially in disease diagnosis or cancer subtypes analysis, multiomics data studies and population stratification [6].

A number of clustering algorithms other have been emerged in the literature. Here we briefly summarize exemplary methods. 1) Algorithms are developed using criteria functions, such as a) sum-of-squared error; b) scattering; c) related minimum variance; d) trace; e) determinant; and, f) invariant criterion [1, 7]; 2) clustering following iterative optimization [8–10]; 3) hierarchical clustering algorithms [11–14]; some conventional hierarchical-based algorithms are, single-linkage [15], complete-linkage [16], median-linkage [17], weighted average linkage [18] and ward linkage [19]. Single linkage (SLink) agglomerative hierarchical approach [15] combines clusters which are nearest to each other and applies Euclidean distance to quantity the nearness between the two neighboring groups. This method is sensitive to the positioning of samples, which sometimes causes an issue of a long chain (called the chaining effect). The hierarchical approach with complete linkage (CLink) [16] tries to reduce the chain effect by constructing groups using farthest-neighbor. However, it is susceptible to outliers. This problem can be overcome by applying average or median distance which was achieved in median linkage (MLink) hierarchical approach [17]. In the hierarchical weighted-average distance linkage (WLink) approach, group sizes are ignored while computing average distances. Consequently, smaller groups get larger weights during clustering [18]. In Ward's linkage (Wa-Link), the clusters are joined based on an optimal value of an objective function. Similarly, in model-based hierarchical clustering [20, 21] an objective function is used. The method presented in [20] is based on the Bayesian analysis and uses multinomial likelihood function and Dirichlet priors. The approach in [21] optimizes the distance between two Gaussian mixture models. 4) Clustering is carried by Bayes classifier [22–26]; 5) by maximum likelihood in an iterative fashion [27–30]. In general, maximum likelihood can be computed via analytical procedure, grid search, hill-climbing procedure or EM algorithm [27, 31–35]; 6) spectral clustering use spectrum of similarity matrix to perform dimensionality reduction before conducting clustering [36], 7) non-negative matrix factorization (NNMF) [37] has also been used for clustering [38–40] and has been useful in handling high-dimensional data; and, 8) support vector clustering (SVC) became popular in recent literature [41–47]. However, its computational complexity is quite high and occasionally it fails to discover meaningful groups [14]. In general, for many applications clustering techniques constructed on maximum likelihood and Bayes approach are still the favored over support vector clustering. Maximum likelihood methods require differential calculus techniques or gradient search to estimate parameters. However, Bayes methods usually require solving complex multi-dimensional integration to reach to the solution. Since Bayes estimation methods has very high computational requirements [1], we prefer maximum likelihood in this paper.

Though many clustering methods have been developed in the literature for various applications [48–54], the problem of achieving a reasonable level of accuracy for high dimensional data still persists. Many of these algorithms fail to perform when the number of features is gradually increased and becomes huge in comparison with the number of samples [55–62]. Many methods that rely on data distribution, suffers from high dimensionality as such case create the problem of singularity of covariance matrix. Therefore, methods based on norm distance (e.g. Euclidean) or centroid based distance prevail in these situations. This is the usual case for many biological applications where generating additional samples is cost prohibitive. In order to deal with the dimensionality issue, in general

either feature transformation or feature selection is applied to reduce (or transform) the data into a parsimonious space before executing clustering operation. This has its own advantages and disadvantages. Inspired by this drawback, we focus on developing a method that can easily and efficiently perform clustering on high dimensional data.

We propose a novel way of handing the data that precedes clustering. A sample (in a vector form) is reformed into a matrix form through a filtering process that simultaneously facilitates more straightforward visualization. This is a critical stage of this concept, as this reformation process can retain a significant amount of useful information for clustering that could otherwise be difficult to capture. Furthermore, we extended EM algorithm to estimate maximum likelihood for samples which appears in the matrix form (i.e. feature matrix) in contrast to the conventional methods which take input samples as feature vectors.

The novel method, which we named 2D–EM, has two steps. The first, filtering part produces a feature matrix for a sample while the subsequent clustering part is based on a modified EM algorithm that is capable of accepting these feature matrices as input. The maximum likelihood estimate via EM algorithm has been modified such that it can consider input as feature matrix instead of feature vector. The details of the method are given in the later section. We observed a significant improvement over many clustering algorithms over a number of transcriptome and methylome datasets evaluated in this study. We first present an overview of the maximum likelihood estimate via EM algorithm and then present our proposed 2D–EM clustering algorithm.

Methods

Overview of maximum likelihood estimate via EM algorithm

Here we briefly present the summary of the maximum likelihood via EM algorithm for clustering [1, 27, 63]. Suppose a d-dimensional sample set is described as $\chi = \{x_1, x_2, \ldots, x_n\}$ with n unlabelled samples. Let number of clusters be defined as c. Let the state of the nature or class label for jth cluster χ_j (for $j = 1, \ldots, c$) be depicted as ω_j. Let $\theta = \{\mu, \Sigma\}$ be any unknown parameter (representing mean μ and covariance Σ). Then the mixture density would be

$$p(x_k|\theta) = \sum_{j=1}^{c} p(x_k|\omega_j, \theta_j)P(\omega_j) \tag{1}$$

where $p(x_k|\ \omega_j, \theta_j)$ is the conditional density, $\theta = \{\theta_j\}$ (for $j = 1 \ldots c$), $x_k \in \chi$ and $P(\omega_j)$ is the a priori probability. The log likelihood can be given by joint density

$$\begin{aligned} L &= \log p(\chi|\theta) = \log \prod_{k=1}^{n} p(x_k|\theta) \\ &= \sum_{k=1}^{n} \log p(x_k|\theta) \end{aligned} \tag{2}$$

If the joint density $p(\chi|\ \theta)$ is differentiable w.r.t. to θ then from Eqs. 1 and 2

$$\nabla_{\theta_i} L = \sum_{k=1}^{n} \frac{1}{p(x_k|\theta)} \nabla_{\theta_i} \left[\sum_{j=1}^{c} p(x_k|\omega_j, \theta_j)P(\omega_j) \right] \tag{3}$$

where $\nabla_{\theta_i} L$ is defined as the gradient of L w.r.t. to θ_i. If θ_i and θ_j are independent parameters and assume a posteriori probability is

$$P(\omega_i, |x_k, \theta) = \frac{p(x_k|\omega_i\theta_i)P(\omega_i)}{p(x_k|\theta)} \tag{4}$$

then from Eq. 4, we can observe that $\frac{1}{p(x_k|\theta)} = \frac{P(\omega_i,|x_k,\theta)}{p(x_k|\omega_i,\theta_i)P(\omega_i)}$. Substituting this value in Eq. 3 and since for any function $f(x)$ its derivative $\partial \log f(x)/\partial x$ can be given as $1/f(x).\ \dot{f}(x)$. We have

$$\nabla_{\theta_i} L = \sum_{k=1}^{n} P(\omega_i|x_k, \theta)\, \nabla_{\theta_i} \log p(x_k|\omega_i, \theta_i) \tag{5}$$

If distribution of the data is normal Gaussian and $\theta_i = \{\mu_i, \Sigma_i\}$ then we can employ Eq. 5 to find E-step and M-step of EM algorithm to find maximum likelihood estimate θ_i. The solution be achieved by.

E-step

$$\phi_{ik} = P(\omega_i|x_k \mu \Sigma)$$

M-step

$$\pi_i = \frac{1}{n} \sum_{k=1}^{n} P(\omega_i|x_k, \mu, \Sigma) \tag{6}$$

$$\mu_i = \frac{\sum_{k=1}^{n} \phi_{ik} x_k}{\sum_{k=1}^{n} \phi_{ik}} \tag{7}$$

$$\Sigma_i = \frac{\sum_{k=1}^{n} \phi_{ik} (x_k - \mu_i)(x_k - \mu_i)^T}{\sum_{k=1}^{n} \phi_{ik}} \tag{8}$$

where π_i is the a priori probability, $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d \times d}$. For a normal distribution case, ϕ_{ik} can be expressed as

$$\begin{aligned} \phi_{ik} &= \frac{p(x_k|\omega_i, \mu_i, \Sigma_i)\pi_i}{\sum_{j=1}^{c} p(x_k|\omega_j, \mu_j, \Sigma_j)\pi_j} \\ &= \frac{|\Sigma_i|^{-1/2} \exp\left[-\frac{1}{2}(x_k - \mu_i)^T \Sigma_i^{-1} (x_k - \mu_i) \right] \pi_i}{\sum_{j=1}^{c} |\Sigma_j|^{-1/2} \exp\left[-\frac{1}{2} \left(x_k - \mu_j \right)^T \Sigma_j^{-1} \left(x_k - \mu_j \right) \right] \pi_j} \end{aligned} \tag{9}$$

For every iteration check whether $L = \sum_{k=1}^{n} \log \sum_{j=1}^{c} \pi_j p(x_k|\omega_j, \mu_j, \Sigma_j)$ is converging. At the convergence of L

this procedure yields maximum likelihood estimate $\hat{\theta}_i = \{\hat{\mu}_i, \hat{\Sigma}_i\}$ (for $i = 1, 2, \ldots, c$).

As it can be observed from the above procedure, the maximum likelihood estimate is possible if the inverse of covariance matrix exists. For high dimensional data (where samples are relatively lower), the computation of maximum likelihood estimate becomes difficult as covariance matrix becomes singular.

2D–EM clustering methodology

In this section, we describe our proposed 2D–EM clustering algorithm. In order to overcome the dimensionality problem, we propose to fold a feature vector $\mathrm{x} \in \mathbb{R}^d$ into a matrix form $X \in \mathbb{R}^{m \times q}$ (where $mq \leq d$, number of rows of a feature matrix X is denoted as m whereas number of columns is denoted as q). Thereafter, we find maximum likelihood estimate using EM algorithm for matrices. The 2D–EM algorithm has two main components: 1) filtering step and 2) clustering step. In the filtering part, a feature vector x is reformed into its matrix form or feature matrix X. In the clustering step, feature matrices (or samples in the form of X) are clustered. Figure 1 illustrates the overall procedure of 2D–EM clustering algorithm.

Input samples are first processed through a filter where each sample is formed as a matrix. Thereafter, these feature matrices are sent to the clustering process.

Here we first describe the clustering part of 2D–EM algorithm for feature matrices to obtain maximum likelihood estimate. Let a sample $X_k \in \mathbb{R}^{m \times q}$ (where $m \leq q$) be formed from $\mathrm{x}_k \in \mathbb{R}^d$ by a filtering process (to be discussed later). We define the mean $M \in \mathbb{R}^{m \times q}$ and covariance $C \in \mathbb{R}^{m \times m}$ for feature matrices.

The class-conditional density for a feature matrix X_k can be described as,

$$p(X_k|\omega_i, \theta_i) = \frac{1}{(2\pi)^{m \times q}|C_i|^{1/2}} \exp\left(-\frac{1}{2} trace\left((X_k - M_i)^T C_i^{-1}(X_k - M_i)\right)\right) \tag{10}$$

The derivative of likelihood function can be obtained in a similar way as that of maximum likelihood estimate and it comes similar to Eq. 5 as

$$\nabla_{\theta_i} L = \sum_{k=1}^{n} P(\omega_i|\mathrm{X}_k, \theta) \nabla_{\theta_i} \log p(\mathrm{X}_k|\omega_i, \theta_i) \tag{11}$$

This fortunately simplifies the derivations of maximum likelihood estimate for feature matrices and the 2D–EM procedure can be described as.

2D E-step

$$\phi_{ik} = P(\omega_i|\mathrm{X}_k, M, C)$$

2D M-step

$$\pi_i = \frac{1}{n}\sum_{k=1}^{n} P(\omega_i|\mathrm{X}_k, M, C) \tag{12}$$

$$M_i = \frac{\sum_{k=1}^{n} \phi_{ik} X_k}{\sum_{k=1}^{n} \phi_{ik}} \tag{13}$$

$$C_i = \frac{\sum_{k=1}^{n} \phi_{ik} (\mathrm{X}_k - M_i)(\mathrm{X}_k - M_i)^T}{\sum_{k=1}^{n} \phi_{ik}} \tag{14}$$

In a similar way, for a normal distribution case, ϕ_{ik} can be expressed as

$$\begin{aligned} \phi_{ik} &= \frac{p(X_k|\omega_i, M_i, C_i)\pi_i}{\sum_{j=1}^{c} p(X_k|\omega_j, M_j, C_j)\pi_j} \\ &= \frac{|C_i|^{-1/2} \exp\left[-\frac{1}{2} trace\left[(\mathrm{X}_k - M_i)^T C_i^{-1}(\mathrm{X}_k - M_i)\right]\right]\pi_i}{\sum_{j=1}^{c} |C_j|^{-1/2} \exp\left[-\frac{1}{2} trace\left[(\mathrm{X}_k - M_j)^T C_j^{-1}(\mathrm{X}_k - M_j)\right]\right]\pi_j} \end{aligned} \tag{15}$$

Again, for every iteration it can be observed if likelihood L is converging.

It can be seen from Eq. 14 that covariance matrix is no longer of $d \times d$ size, however, it is reduced to size $m \times m$. Since $m^2 \leq d$, theoretically we can say that the size of covariance matrix is reduced to the square root (or less) of the data dimensionality. This reduction is achieved without performing linear or non-linear transformation (of data). Furthermore, this enables us to use Eq. 15 effectively as

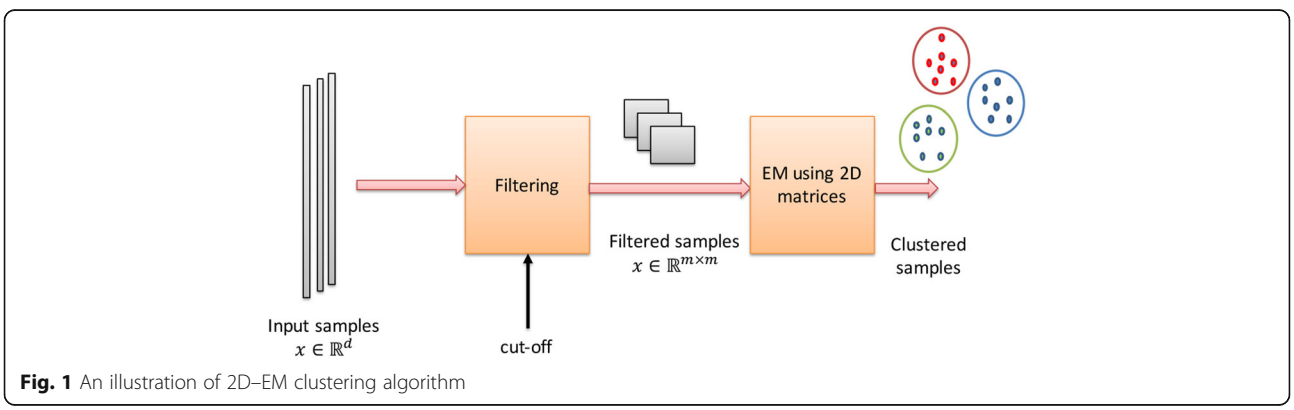

Fig. 1 An illustration of 2D–EM clustering algorithm

singularity problem of C_i matrix is reduced at least by the square root of the data dimensionality.

Next, we discuss the filtering process. The objective of this process is to form a sample $x \in \mathbb{R}^d$ into a matrix $X \in \mathbb{R}^{m \times q}$ form. For convenience, here we use $q = m$; i.e., size of X would be $m \times m$. This filtering process has two parts: 1) feature selection, and 2) matrix arrangement.

In the feature selection part, we perform ANOVA to find p-values for each of the features and then retain the top m^2 features. Here we have used p-values as a prototype to filter genes or features. However, one can use any other scheme, e.g. regression methods (logistic regression, linear regression, Poisson regression, Lasso etc.) depending upon the application or specific type of data used. Since we do not know the class labels of data, we need to find temporary class labels to compute p-values for features. Therefore, to obtain p-values, we perform hierarchical clustering to find c clusters. Thereafter, from the known labels we can compute p-values which will help us to remove some features. This process will give us a feature vector $y \in \mathbb{R}^{m^2}$ where $m^2 \leq d$ and features in y is arranged corresponding to the low to high p-values.

In the matrix arrangement part, we arrange y to get a feature matrix $X \in \mathbb{R}^{m \times m}$. To arrange features in X systematically so that any two samples can be compared without having a conflict, we applied a simple rule. We computed the mean μ_y from all y samples and then arranged features of μ_y in ascending order. Thereafter, we arranged features of y corresponding to the order of features of μ_y. This allows us to put features in a common format for all the samples. Next, we reshape $y \in \mathbb{R}^{m^2}$ so that it becomes $X \in \mathbb{R}^{m \times m}$.

The value of m can be computed as follows. First, the cut-off for p-values will reduce dimensions from d to h (where $h \leq d$). Then m can obtained as $m = \lfloor \sqrt{h} \rfloor$, where $\lfloor \sqrt{h} \rfloor \leq \sqrt{h}$ and $[\cdot]$ is an integer; i.e., m is an integer smaller or equal to $\sqrt{h}$. The arrangement of feature matrix process is summarized in Table 1. The filtering process is summarized in Table 1.

It is also possible to visualize feature matrix X and can be compared with other samples to see the difference or similarity. Figure 2 provides an illustration of visualization of high dimensional data. A feature vector $x \in \mathbb{R}^d$ is constructed as a feature matrix $X \in \mathbb{R}^{m \times q}$ through the filtering process (as described in Table 1). For this illustration, two different groups of samples (Type-A and Type-B) which were difficult to visualize in $\mathbb{R}^d$ space, are shown on $\mathbb{R}^{m \times q}$ space. The visualization of feature matrix is more meaningful in the matrix space.

To further demonstrate this with transcriptome data, we consider six samples from ALL dataset (data used in this paper are described later in Section 3.1). These samples are randomly picked for this illustration. Three samples belong to cluster *acute lymphoblastic leukemia* (ALL) and the other three samples belong to cluster *acute myeloid leukemia* (AML). The number of features (or dimensions) of these samples is 7129 and it is impossible to visualize data in 7129-dimensional space. However, using filtering (from Table 1) we can visualize each sample as a matrix (see Fig. 3). Just by looking at the patterns of these feature matrices, it can be observed that samples from ALL are different from that of AML. The patterns of AML feature matrices have high intensity (or shades) at specific locations compared to the patterns of ALL feature matrices. This reformation of sample from vector to matrix form assist in data visualization and pattern recognition. Similarly, it would also improve the power of detection for a clustering method provided if the method was designed well to utilize this information.

Table 1 Arrangement of features into $m \times m$ matrix

Feature Selection

1. Given $x \in \chi$ in a d-dimensional space.
2. Perform hierarchical clustering on all samples x to find temporary class labels.
3. Using these class labels find p-values for all the d features.
4. Find m by placing a threshold or cut-off on p-values (e.g. cut-off for p-values could be 0.01).
5. Retaining the top m^2 features will give us a sample $y \in \mathbb{R}^{m^2}$, where all y samples form a sample set $Y \in \mathbb{R}^{m^2 \times n}$.

Matrix arrangement

6. Compute mean $\mu_y = \frac{1}{n} \sum_{y \in Y} y$.
7. Arrange features of μ_y in ascending order and note the indices.
8. Arrange features of y by following the indices from step 7.
9. Reshape a sample y to a matrix $X \in \mathbb{R}^{m \times m}$.

Results and discussion

In order to verify the performance of 2D–EM clustering algorithm, we employed 6 transcriptome and 2 methylome datasets described below. We used several clustering algorithms and employed Rand score [64] and adjusted Rand index [65] as a performance measure to compare the clustering algorithms in this study. The Rand scoring reflects how well the group labels were reproduced using unlabeled data, and a high score build confidence in the methods ability to detect novel groups in novel data for which no phenotype labels are available. These are well known measures to gauge the performance of clustering algorithm [66]. The results are described in the 'Clustering on transcriptome data' and 'Clustering on methylome data' sections.

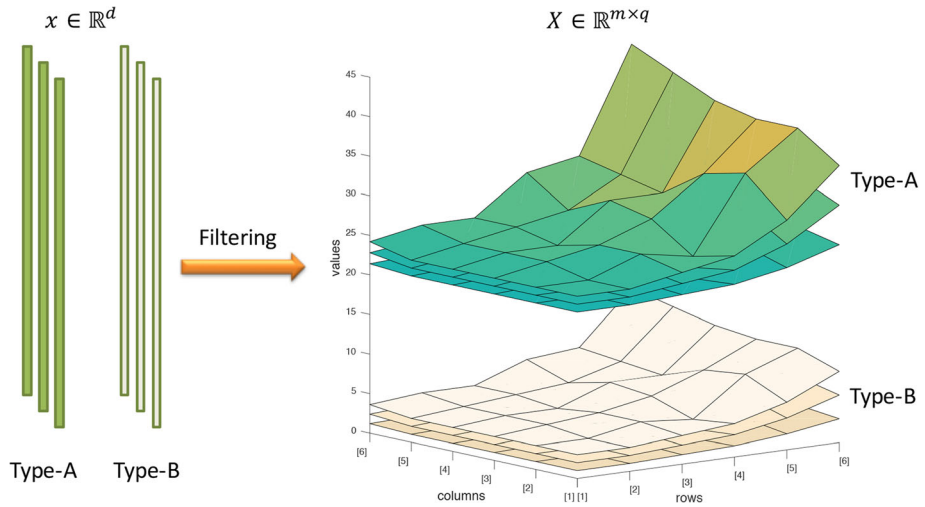

Fig. 2 Visualization of high dimensional data

Biomedical datasets

Acute leukemia dataset [67]: contains DNA microarray gene expressions of acute leukemia samples. Two kinds of leukemia are provided, namely acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). It consists of 25 AML and 47 ALL bone marrow samples over 7129 probes. The features are all numeric having 7129 dimensions.

Small round blue-cell tumor (SRBCT) dataset [68]: has 83 samples of the RNA expression profiles of 2308

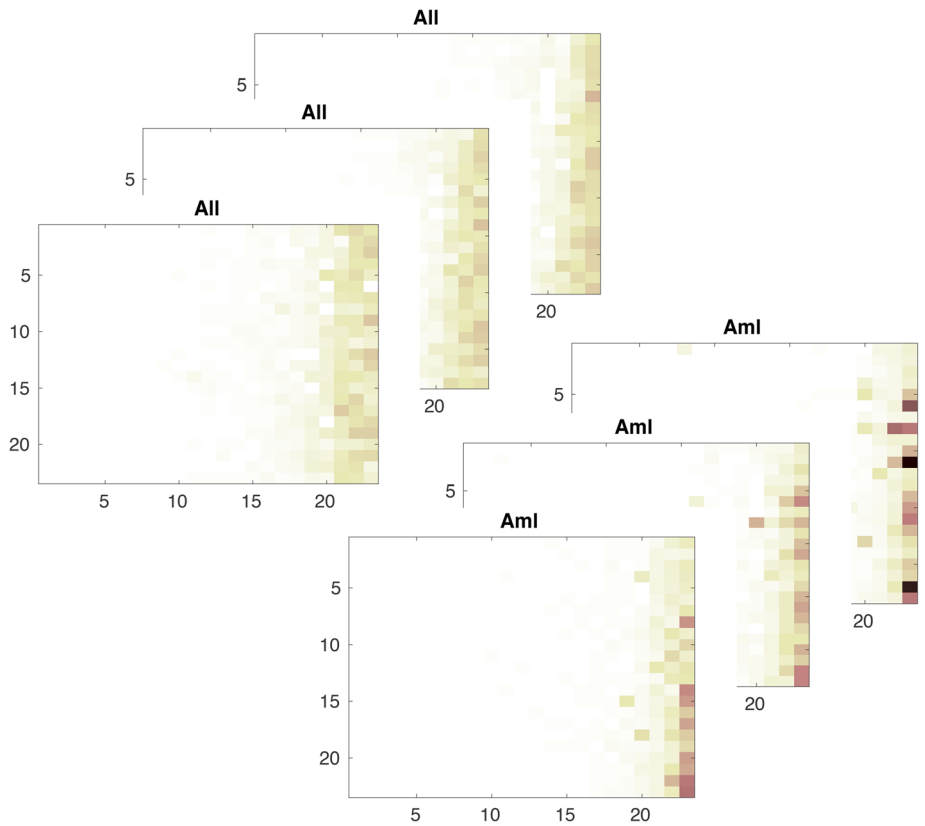

Fig. 3 Visualization of feature matrix: *acute lymphoblastic leukemia* (ALL) vs. *myeloid leukemia* (AML). An ALL sample or feature vector $x \in \mathbb{R}^d$ is transformed to feature matrix $X \in \mathbb{R}^{m \times m}$ using the procedure outlined in Table 1. These feature matrices are shown at top right side of the figure. Similarly, a sample of AML is also transformed to feature matrix and shown at bottom right side of the figure

Table 2 Transcriptome and methylome datasets

Datasets	Features	Samples	Classes
ALL Leukemia	7129	72	2
SRBCT	2308	83	4
MLL	12,582	72	3
ALL Subtype	12,558	327	7
GCM	16,063	198	14
Lung Cancer	12,553	181	2
Gastric Cancer	27,579	64	2
Hepatocellular Carcinoma	27,579	40	2

genes. The tumors are the Ewing family of tumors (EWS), Burkitt lymphoma (BL), neuroblastoma (NB), and rhabdomyosarcoma (RMS). The dataset consists of 29, 11, 25 and 18 samples of EWS, BL, RMS and NB, respectively.

MLL Leukemia [69]: has three groups ALL, AML leukemia and mixed lineage leukemia (MLL). The dataset contains 20 MLL, 24 ALL and 28 AML. The dimensionality is 12,582.

ALL subtype dataset [70]: contains 12,558 gene expressions of acute lymphoblastic leukemia subtypes. It has 7 groups namely E2A-PBX1, BCR-ABL, MLL, hyperdiploid >50 chromosomes ALL, TEL-AML1, T-ALL and other (contains diagnostic samples that did not fit into any of the former six classes). Samples per group are 27, 15, 20, 64, 79, 43 and 79, respectively.

Global cancer map (GCM) [71]: has 190 samples over 14 classes with 16,063 gene expressions.

Lung Cancer [72]: contains gene expression levels of adenocarcinoma (ADCA) and malignant mesothelioma (MPM) of the lung. In total, 181 tissue samples with 12,533 genes are given where 150 belongs to ADCA and 31 belongs to MPM.

Table 3 Rand score (highest values are highlighted as bold faces)

Method	SRBCT	ALL	MLL	ALL subtype	GCM	Lung cancer
K-means	0.58	0.53	0.78	0.64	0.84	0.72
CLink	0.30	0.49	0.54	0.52	0.71	0.70
ALInk	0.30	0.56	0.35	0.51	0.38	0.71
Ward-Link	0.44	0.56	0.78	0.53	0.84	0.80
Weighted-Link	0.30	0.52	0.51	0.52	0.61	0.71
Mlink	0.30	0.55	0.35	0.48	0.54	0.71
Spectral Clustering	0.39	0.51	0.56	0.63	0.55	0.71
NNMF Clustering	**0.66**	0.50	0.74	0.64	0.83	0.63
Mclust	0.51	0.50	0.61	0.30	0.83	0.57
2D–EM	0.65	**0.62**	**0.80**	**0.78**	**0.87**	**0.84**

Table 4 Adjusted Rand index (highest values are highlighted as bold faces)

Method	SRBCT	ALL	MLL	ALL subtype	GCM	Lung cancer
Kmeans	0.13	0.03	0.47	0.15	0.19	0.22
CLink	0.00	−0.03	0.13	0.00	0.09	−0.02
ALInk	0.00	0.05	0.00	−0.01	0.01	−0.01
Wa-Link	0.00	0.09	0.51	0.00	0.17	0.41
Wt-Link	0.00	−0.03	0.08	0.00	0.07	−0.01
Mlink	0.00	0.02	0.00	−0.01	0.08	−0.01
Spectral Clustering	−0.02	0.02	0.02	0.00	0.07	−0.01
NNMF Clustering	0.18	0.00	0.42	0.11	0.17	0.26
Mclust	−0.02	−0.01	0.21	−0.01	0.09	0.05
2D–EM	**0.19**	**0.23**	**0.57**	**0.26**	**0.22**	**0.62**

Gastric Cancer [73]: 32 pairs of gastric cancer and normal (adjacent) tissue were profiled using Illumina Infinium HumanMethylation27 BeadChip. 27,579 CpG sites were interrogated at a single-nucleotide resolution. Both Beta- and M-values statistics were calculated from the methylated and unmethylated signals as described in [74].

Hepatocellular Carcinoma [75]: 20 pairs of hepatocellular tumor and their non-tumor tissue counterparts were evaluated using the same platform (27,579 CpG sites) and processed in the same manner as in Gastric cancer dataset.

A summary of the transcriptome and methylome datasets is depicted in Table 2. It is evident from the table that the number of features (genes or CpG site methylation state) is much larger than the number of samples for all the datasets. This creates SSS problem in all the cases.

Clustering on transcriptome data

In this subsection, we show the performance of various clustering methods in terms of Rand score [64] over 6 transcriptome datasets. Rand score shown here represents an average taken from over 10 repetitions. Rand score is similar to clustering accuracy and it value lies between 0 and 1. We also used adjusted Rand index [65], which assumes the generalized hypergeometric model. Adjusted Rand index can attain wider range of values than Rand score.

Table 5 Percentage improvement of 2D–EM clustering method over other existing clustering methods

Parameter	SRBCT	ALL	MLL	ALL subtype	GCM	Lung cancer
Rand Score	−1.5	10.7	2.6	21.9	3.6	5.0
Adjusted Rand Index	5.6	155.6	11.8	73.3	21.1	51.2

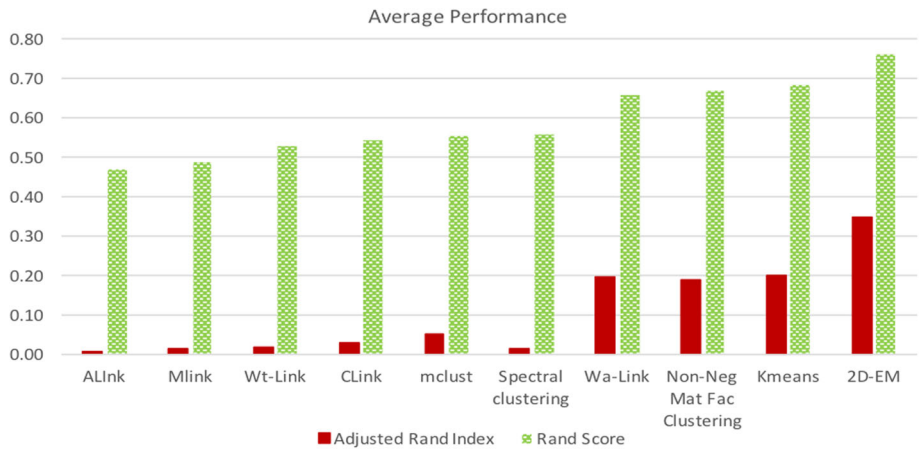

Fig. 4 Comparison of average performance (in terms of Rand score and Adjusted Rand index)

Rand and adjusted Rand scores

For 2D–EM clustering algorithm we use 0.01 as a cut-off during the filtering process (the reasoning behind selecting this particular cut-off is described in section 'Effect of using filter'). Table 3 depicts the Rand score analysis and Table 4 shows adjusted Rand index. We have employed several clustering methods for comparison. These methods are k-means, hierarchical clustering methods (SLink, CLink, ALink, MLink, Ward-Link and Weighted-Link), spectral clustering, mclust [76] and NNMF clustering. For k-means and hierarchical clustering methods, packages from MATLAB software were used. For NNMF clustering method, package provided by ref. [38] was used. For spectral clustering, package provided by ref. [77] was used. In all the cases, only data was provided with the number of cluster information.

It can be observed from Table 3 that for SRBCT dataset, NNMF clustering is showing 0.66 Rand score followed by 0.65 of 2D–EM. However, adjusted Rand index (Table 4) for SRBCT is better for 2D–EM. For all other datasets 2D–EM is performing the best in terms of Rand score and adjusted Rand index (Table 3 and Table 4).

For an instance, we can observe that from Table 3, 2D-EM scored highest Rand score of 0.62 followed by ALink (0.56) and Ward-link (0.56) on ALL dataset. For MLL k-means and Ward-link scored 0.78 and 2D–EM was able to score 0.80. In the case of ALL subtype, 2D–EM scored 0.78 followed by k-means (0.64) and NNMF (0.64). For GCM, 2D–EM got 0.87 followed by k-means (0.84) and Ward-link (0.84). For Lung Cancer, Ward-link scored 0.80 and 2D–EM reached 0.84. We can also observe that spectral clustering underperforming when the dimensionality is large. Similarly, many clustering methods (not reported here) did not provide results due to high number of features.

Similarly, we can see from Table 4 that 2D–EM is way ahead on ALL dataset by attaining 0.23 adjusted Rand index followed by second best of 0.09 by Ward-link. For MLL dataset, 2D–EM scored 0.57 followed by Ward-link (0.51) and mclust (0.51). In case of ALL subtype and GCM datasets, 2D–EM (0.26, 0.22) is followed by k-means (0.15, 0.19). For Lung dataset, 2D–EM scored 0.62 followed by mclust (0.36).

The improvement (in terms of Rand score and adjusted Rand index) of 2D–EM over the best performing existing method has been depicted in Table 5. It can be noticed that the best percentage improvement for Rand score compared to the best performing clustering method is 21.9%. Similarly, the best percent improvement in terms of adjusted Rand index is 155.6%.

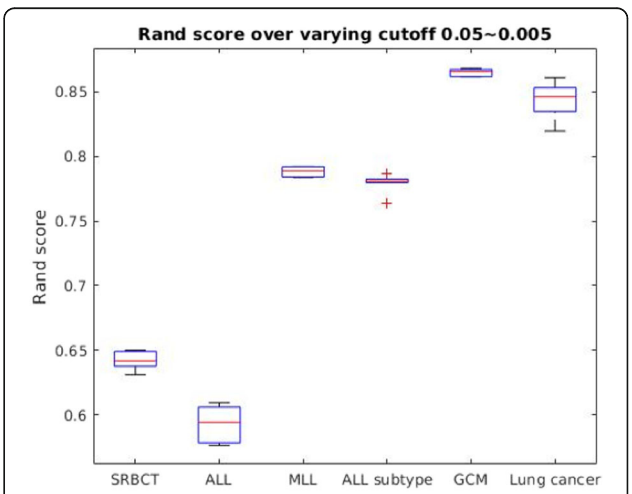

Fig. 5 Box plot showing the effect of changing cut-off value for 2D–EM clustering algorithm

Average performance

We have also compared the average of Rand score and adjusted Rand index over all the datasets used. The comparison is depicted in Fig. 4. The comparison of average performance is interesting. It can be seen that k-means clustering algorithm performs quite reasonably for high dimensional data. Several clustering algorithms have been proposed after k-means algorithm, yet for high dimensional data the average performance has not been improved. Apart from k-means algorithm, Ward-Link hierarchical clustering, NNMF clustering, mclust and spectral clustering were able to attain reasonable level of performance. The 2D–EM clustering algorithm was able to attain 11.4% improvement on Rand score, and 75.0% improvement on adjusted Rand index over the best performing method. Therefore, it can be concluded that in all the cases 2D–EM was able to achieve very promising results.

Effect of using filter

The 2D–EM clustering algorithm uses a filtering step to arrange a feature vector into a feature matrix. We want to analyze the effect of applying this filter to other clustering algorithms. In order to perform this analysis, we preprocess data to retain top m^2 features by filtering before executing other clustering algorithms (note samples are not reshaped in matrix form for other methods as this would require changing the mathematics of

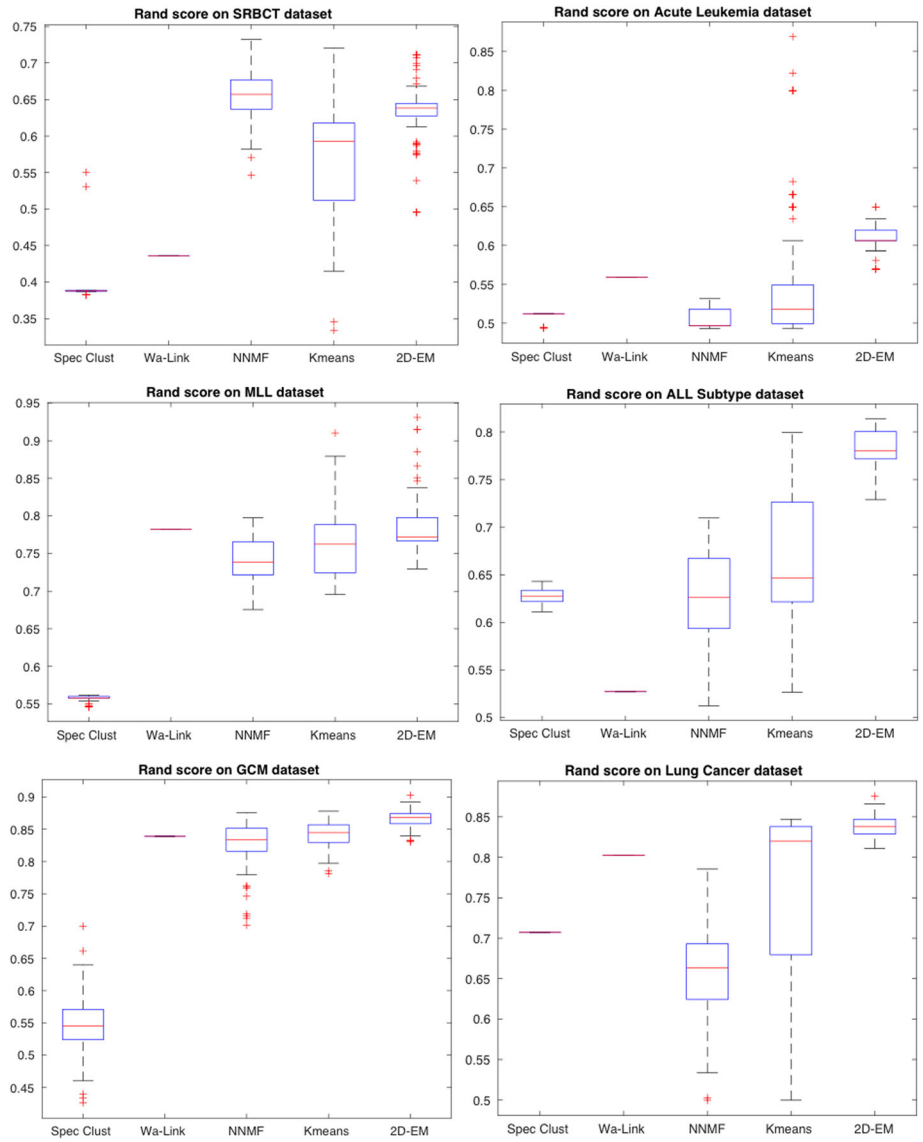

Fig. 6 Rand score of five best performing methods over 100 runs

algorithms). The detailed results are given in Additional file 1: it can be observed from Tables S1, S2, S3 and S4 that after applying filter for other clustering methods, the performance doesn't improve significantly. Therefore, the evidence of bias due to filtering process is weak.

Effect of variable cut-off

In order to illustrate the effect of changing the cut-off value for the 2D–EM clustering algorithm, we varied cut-off value from 0.05 to 0.005 and noted the Rand score over 10 repetitions. The box-plot with the corresponding results is shown in Fig. 5. It can be noticed from Fig. 5, that varying cut-off value over a range (0.05~0.005) does not significantly change the Rand score of the algorithm. Therefore, the selection of 0.01 cut-off value in the previous experiment is not a sensitive choice.

Clock time

The processing (clock) time of 2D–EM clustering algorithm when run on Linux platform (Ubuntu 14.04 LTS, 64 bits) having 6 processors (Intel Xeon R CPU E5–1660 v2 @ 3.70GHz) and 128 GB memory per repetition is as follows. On SRBCT dataset, 2D–EM clustering algorithm took 11.4 s. Similarly, on ALL, MLL, ALL subtype, GCM and Lung datasets, processing time were 8.7 s, 47.1 s, 286.5 s, 358.2 s and 82.0 s, respectively.

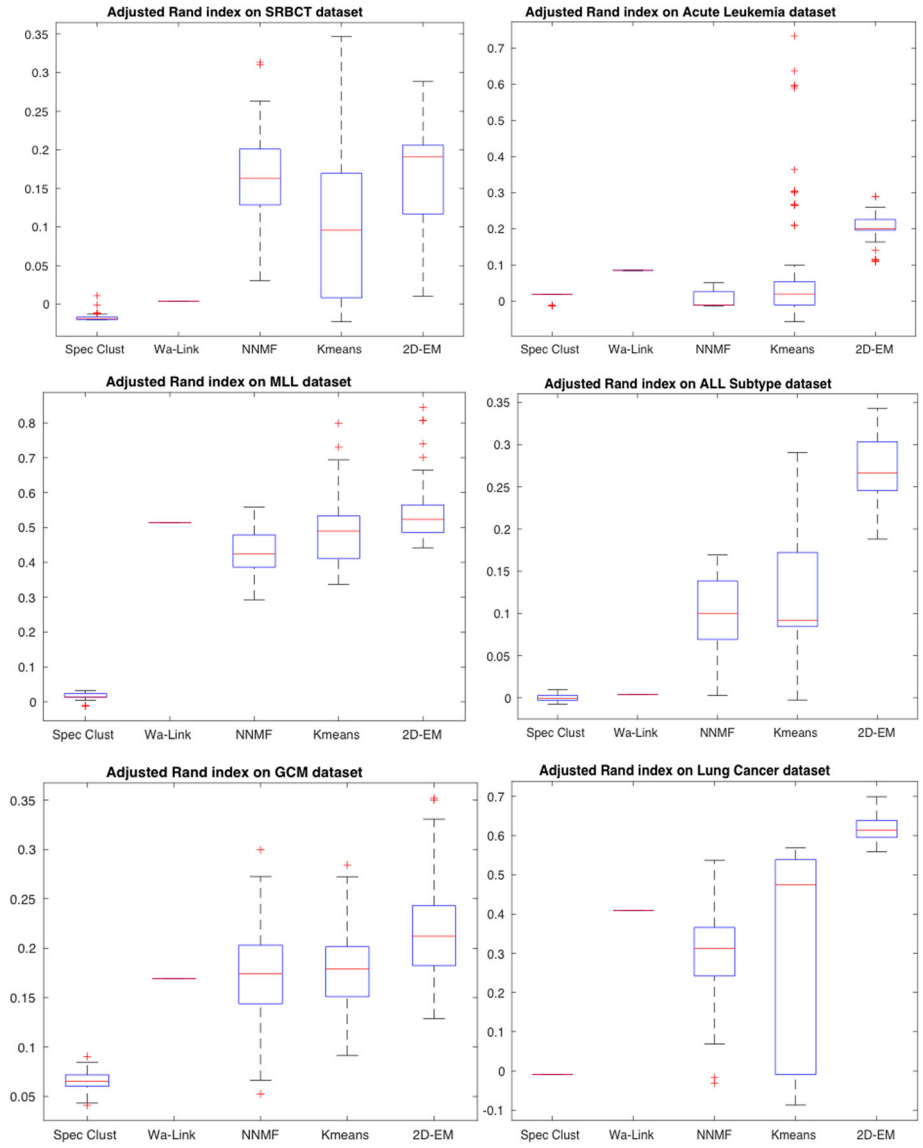

Fig. 7 Adjusted Rand index over 100 runs

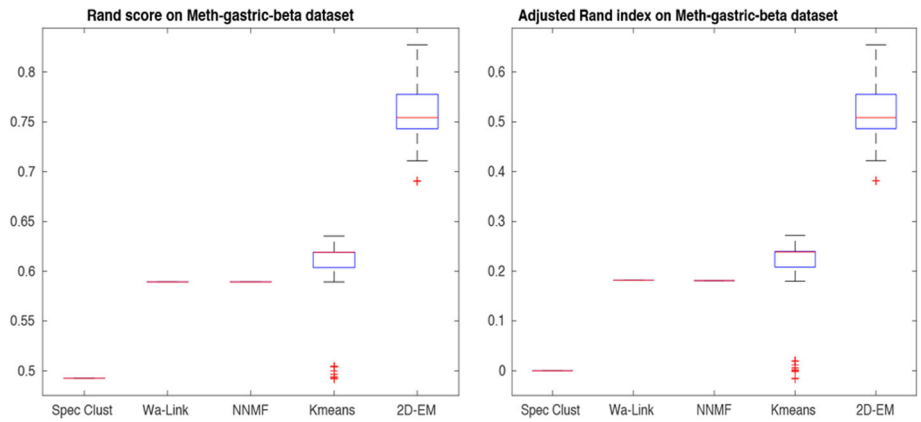

Fig. 8 Rand score and adjusted Rand index on Gastric cancer methylation data (Beta-values)

Therefore, for all the transcriptome datasets used in this study, the processing time for 2D–EM clustering algorithm was within 6 mins.

Consistency

To verify the consistency or stability of 2D–EM clustering algorithm, we employed top five performing clustering algorithms and obtain boxplots of Rand score and adjusted Rand index over all the transcriptome datasets used. The results are derived from over 100 runs. Figure 6 depicts boxplot of Rand score of 5 best methods (spectral clustering, Wa-Link, NNMF, k-means and 2D–EM). It can be observed that on SRBCT dataset NNMF is showing superior performance followed by 2D–EM clustering algorithm. However, on all the remaining 5 datasets (ALL, MLL, ALL Subtype, GCM and Lung Cancer), 2D–EM is outperforming all the clustering methods. Similarly, adjusted Rand index was computed on the same datasets and shown in Fig. 7. Again, 2D–EM clustering methodology outperformed all the clustering methods in terms of adjusted Rand index.

Clustering on methylome data

To show the utility of 2D–EM methodology we evaluated two additional datasets of clinical relevance. While in previous examples we showed commonly used transcriptome data, the full understanding of biological phenomena can only be achieved by considering multiple genomics 'layers'. To this end, we compared the Rand score and adjusted Rand index on DNA methylation data. Epigenetic modifications measured in those datasets are known to affect a wide range of biological processes and diseases phenotypes [78]. As we are approaching the era of personalized medicine, clustering of different genomic components will continue to rise in prominence.

For this purpose, we compared the performance of the best 5 methods (selected based on performance with transcriptome data). These methods are spectral clustering,

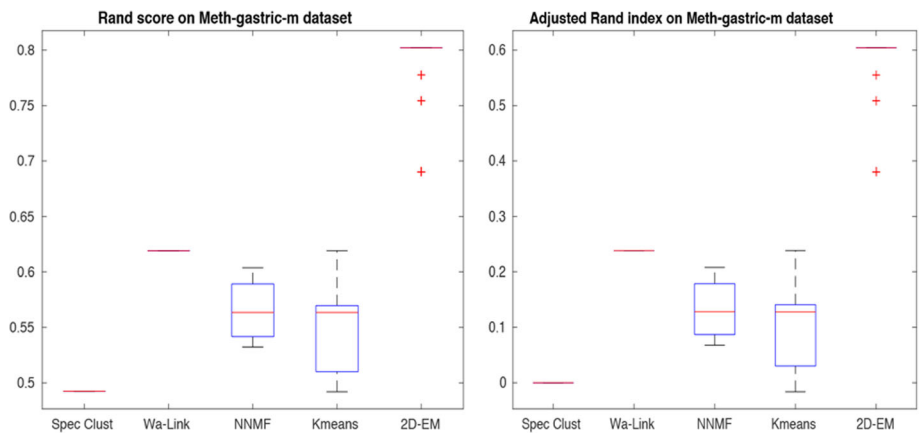

Fig. 9 Rand score and adjusted Rand index on Gastric cancer methylation data (M-values)

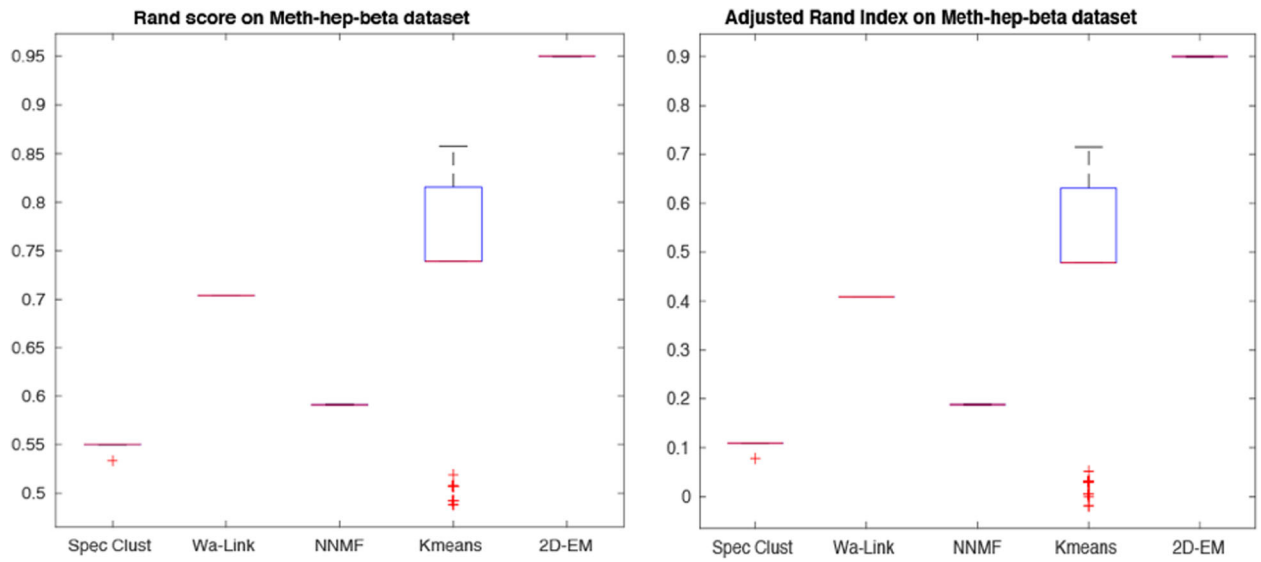

Fig. 10 Rand score and adjusted Rand index on Hepatocellular carcinoma (Beta-values)

Ward-link hierarchical clustering, NNMF, k-means and 2D–EM. Figure 8 depicts Rand score and adjusted Rand index on Gastric cancer methylation data using Beta-values over 100 runs. It can be clearly observed that 2D–EM is outperforming other methods even when different type of data is tested. Similarly, Fig. 9 shows the results on Gastric data using alternative M-values, again for over 100 runs. Again 2D–EM accurately recreated the phenotype labels.

We have also carried out tests on Hepatocellular carcinoma data, with results shown in Figs. 10 and 11 for Beta- M-values respectively. Similar to the Gastric dataset, 2D–EM is achieving very promising results for both Beta- and M-values.

Conclusions

By looking at the nature of data readily found biological sciences, in this work we proposed 2D–EM clustering algorithm. This methodology clusters a given data in two steps. In the first step, it reformats a feature vector to a matrix form and, in the second part, it conducts the clustering. The advantage of 2D–EM algorithm is that it

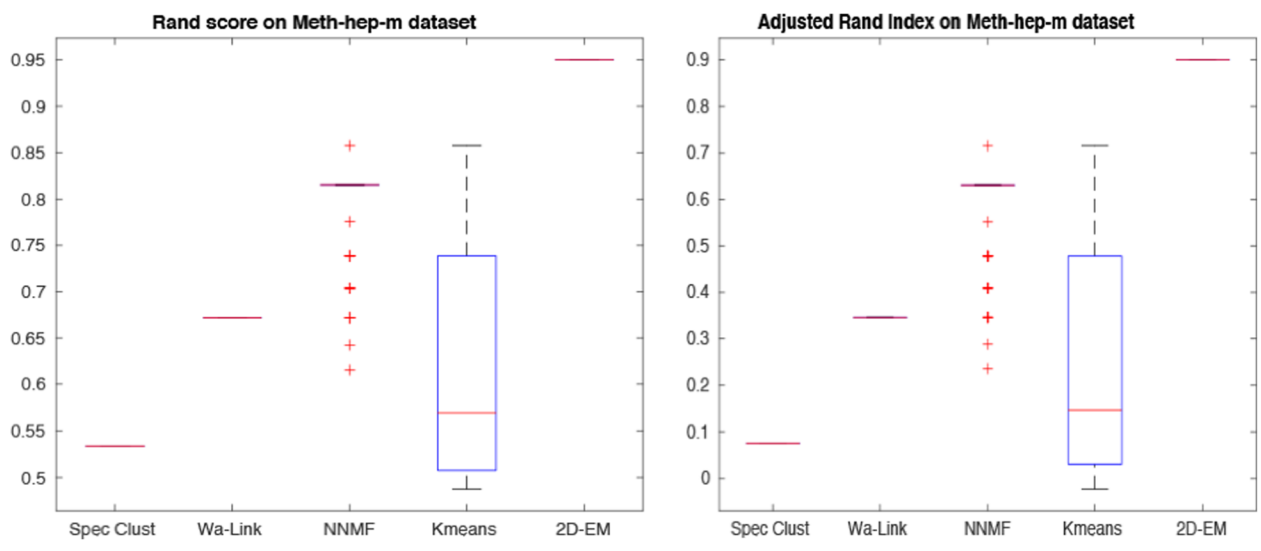

Fig. 11 Rand score and adjusted Rand index on Hepatocellular carcinoma (M-values)

can perform clustering at high dimensional space (compared to the number of samples) by effectively incorporating data distribution information via its covariance matrix. The proposed method avoids the singularity issue by folding a feature vector into a feature matrix. This reduces the dimensionality from d to less than $\sqrt{d}$. Thereby, distribution information along with distance information can be used to cluster a sample. The algorithm was compared to several existing clustering algorithms over a number of transcriptome and methylome datasets, and managed to accurately reproduce the phenotype labels that were hidden from the analysis. MATLAB package of 2D–EM clustering algorithm can be found by visiting our website (http://www.riken.jp/en/research/labs/ims/med_sci_math or http://www.alok-ai-lab.com). In the future, we will investigate ways to extend the present method to Bayesian estimation and hierarchical methods.

Additional file

Additional file 1: In this file the bias of using filtering process is analyzed. Here, we analyzed the effect of applying the filter (which was used for 2D–EM algorithm) to other clustering algorithms. We preprocess data to retain top m^2 features. The m^2 values for all datasets at 0.01 cut-off were as follows: 1156 (SRBCT), 529 (ALL), 6084 (MLL), 1444 (ALL subtype), 15,129 (GCM) and 5625 (Lung Cancer). Then clustering algorithms are applied to see the difference in performance (both in Rand score and adjusted Rand index). **Table S1** and **Table S2** show the Rand score and adjusted Rand score when filtering step is applied. **Table S3** and **Table S4** show the variations in Rand score and adjusted Rand score after filtering compared to before filtering process. (DOCX 25 kb)

Abbreviations

2D: Two-dimension; ADCA: Adenocarcinoma; ALL: Acute lymphoblastic leukemia; AML: Acute myeloid leukemia; BL: Burkitt lymphoma; CLink: Complete Linkage; EM: Expectation-maximization; EWS: Ewing family of tumors; GCM: Global Cancer Map; MLink: Median Linkage; MLL: Mixed lineage leukemia; MPM: Malignant mesothelioma; NB: Neuroblastoma; NNMF: Non-negative matrix factorization; RMS: rhabdomyosarcoma; SLink: Single linkage; SSS: Small sample size; SVC: Support vector clustering; Wa-Link: Ward's linkage; WLink: Weighted average distance linkage

Acknowledgments

Not applicable

Funding

This study has been supported by the JST CREST (JPMJCR1412). PJK was supported by JSPS Postdoctoral Fellowship (15F15776). The publication charges of this article were funded by JST CREST grant JPMJCR1412.

Availability of data and materials

The datasets used and analysed during the current study are publically available online.

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 18 Supplement 16, 2017: 16th International Conference on Bioinformatics (InCoB 2017): Bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-16.

Authors' contributions

AS conceived and implemented the algorithm and performed analysis and experiments. PJK processed and analysed the methylome data and calculated the corresponding Beta- and M-values. AS and PJK wrote the manuscript. TT supervised the project. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Center for Integrative Medical Sciences, RIKEN Yokohama, Yokohama 230-0045, Japan. [2]CREST, JST, Yokohama 230-0045, Japan. [3]Institute for Integrated and Intelligent Systems, Griffith University, 170 Kessels Rd, Nathan, QLD 4111, Australia. [4]Medical Research Institute, Tokyo Medical and Dental University, Tokyo 113-8510, Japan. [5]School of Engineering and Physics, University of the South Pacific, Laucala Bay Rd, Suva, Fiji.

Published: 28 December 2017

References

1. Duda RO, Hart PE, Stork DG: Pattern Classification, 2nd edn. New York: Wiley-Interscience; 2000.
2. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010;31(8):651–66.
3. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013;110(11):4245–50.
4. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a Resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118.
5. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572–3.
6. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016;48(5):500.
7. Maimon O, Rokach L: Data mining and knowledge discovery handbook, 2nd edn: Springer-Verlag New York incorporated; 2010.
8. Fisher D. Iterative optimization and simplification of hierarchical clusterings. J Artif Intell Res. 1996;4(1):147–79.
9. Dhillon IS, Guan Y, Kogan J. Iterative clustering of high dimensional text data augmented by local search. In: The 2002 IEEE international conference on data mining; 2002. p. 131–8.
10. Fayyad UM, Reina CA, Bradley PS. Initialization of iterative refinement clustering algorithms. In: Proceedings of the 4th international conference on Knowledge Discovery & Data Mining (KDD98). Menlo Park, California: AAAI Press; 1998. p. 194–8.
11. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Second ed. New York: Springer; 2009.
12. Heller KA, Ghahramani Z. Bayesian hierarchical clustering. In: Twenty-second international conference on machine learning (ICML); 2005.
13. Farrell S, Ludwig C. Bayesian and maximum likelihood estimation of hierarchical response time models. Psychon Bull Rev. 2008;15(6):1209–17.
14. Sharma A, Boroevich K, Shigemizu D, Kamatani Y, Kubo M, Tsunoda T. Hierarchical maximum likelihood clustering approach. IEEE Trans Biomed Eng. 2017;64(1):112–22.
15. Sibson R. SLINK: an optimally efficient algorithm for the single-link cluster method. Comput J (British Computer Society). 1973;16(1):30–4.
16. Defays D. An efficient algorithm for a complete link method. Comput J (British Computer Society). 1977;20(4):364–6.

17. Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis, 5th edn. Chichester: Wiley Series in Probability and Statistics; 2011.
18. Podani J. Multivariate data analysis in ecology and systematics: a methodological guide to the SYN-TAX 5.0 package. Amsterdam: SPB Academic Publishing. 1994. ISBN: 9051030940.
19. de Amorim RC. Feature relevance in Ward's hierarchical clustering using the L (p) norm. J Classif. 2015;32(1):46–62.
20. Vaithyanathan S, Dom B. Model-based hierarchical clustering. In: Proceedings of 16th conference uncertainty in artificial intelligence; 2000. p. 599–608.
21. Goldberger J, Roweis S. Hierarchical clustering of a mixture model. NIPS. 2005:505–12.
22. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013; 29(20):2610–6.
23. Liu JS, Zhang JL, Palumbo MJ, Lawrence CE. Bayesian clustering with variable and transformation selections. Bayesian Statistics. 2003;7:249–75.
24. Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE Jr. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet. 2006;7(2):295–302.
25. Chen C, Durand E, Forbes F, François O. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes. 2007;7(5):747–56.
26. Ramoni M, Sebastiani P, Cohen P. Bayesian clustering by dynamics. Mach Learn. 2002;47(1):91–121.
27. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc. 1977;39(1):1–38.
28. Misztal I. Comparison of computing properties of derivative and derivative-free algorithms in variance-component estimation by REML. J Anim Breed Genet. 1994;111(1–6):346–55.
29. Denoeux T. Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans Knowl Data Eng. 2013;25(1):119–30.
30. Sharma A, Shigemizu D, Boroevich KA, Lopez Y, Kamatani Y, Kubo M, Tsunoda T. Stepwise iterative maximum likelihood clustering approach. Bmc Bioinformatics. 2016;17(319):1–14.
31. Davidon WC. Variable metric method for minimization. SIAM J Optim. 1991; 1(1):1–17.
32. Adachi J, Hasegawa M: MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. 1996.
33. Long JS. Regression models for categorical and limited dependent variables. London: Sage Publications; 1997.
34. Felsenstein J, Churchill GA. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol. 1996;13(1):93–104.
35. Fletcher R, Powell MJD. A rapidly convergent descent method for minimization. Comput J. 1963;6(2):163–8.
36. von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4): 395–416.
37. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
38. Li Y, Ngom A. The non-negative matrix factorization toolbox for biological data mining. Source Code Biol Med. 2013;8(1):10.
39. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004; 101(12):4164–9.
40. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007;23(12):1495–502.
41. Chiang J-H, Hao P-Y. A new kernel-based fuzzy clustering approach: support vector clustering with cell growing. IEEE Trans Fuzzy Syst. 2003;11(4):518–27.
42. Lee J, Lee D. An improved cluster labeling method for support vector clustering. IEEE Trans Pattern Anal Mach Intell. 2005;27(3):461–4.
43. Lee J, Lee D. Dynamic characterization of cluster structures for robust and inductive support vector clustering. IEEE Trans Pattern Anal Mach Intell. 2006;28(11):1869–74.
44. Horng S-J, Su M-Y, Chen Y-H, Kao T-W, Chen R-J, Lai J-L, Perkasa CD. A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Syst Appl. 2011;38(1):306–13.
45. Jun S, Park S-S, Jang D-S. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst Appl. 2014;41(7):3204–12.
46. Wang K, Liang C, Liu J, Xiao H, Huang S, Xu J, Li F. Prediction of piRNAs using transposon interaction and a support vector machine. BMC Bioinformatics. 2014;15(1):419.
47. Ben-Hur A, Horn D, Siegelmann HT, Vapnik V. Support vector clustering. J Mach Learn Res. 2001;2:125–37.
48. Guha S, Mishra N. Clustering data streams. In: Garofalakis M, Gehrke J, Rastogi R, editors. Data stream management: processing high-speed data streams. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 169–87.
49. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: The IEEE conference on computer vision and pattern recognition, vol. 2015.
50. Wang YX, Xu H. Noisy sparse subspace clustering. J Mach Learn Res. 2016; 17:1–41.
51. Cohen MB, Elder S, Musco C, Musco C, Persu M. Dimensionality reduction for k-means clustering and low rank approximation. In: Proceedings of the forty-seventh annual ACM symposium on theory of computing. Portland, Oregon, USA: 2746569: ACM; 2015. p. 163–72.
52. Esmin AAA, Coelho RA, Matwin S. A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artif Intell Rev. 2015;44(1):23–45.
53. Chi EC, Lange K. Splitting methods for convex clustering. J Comput Graph Stat. 2015;24(4):994–1013.
54. Liang XL, Li WF, Zhang Y, Zhou MC. An adaptive particle swarm optimization method based on clustering. Soft Comput. 2015;19(2):431–48.
55. Sharma A, Imoto S, Miyano S. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans Comput Biol Bioinform. 2012;9(3):754–64.
56. Sharma A, Paliwal KK. A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. Pattern Recogn. 2012;45(6):2205–13.
57. Paliwal KK, Sharma A. Improved direct LDA and its application to DNA microarray gene expression data. Pattern Recogn Lett. 2010;31(16):2489–92.
58. Sharma A, Paliwal KK. Regularisation of eigenfeatures by extrapolation of scatter-matrix in face-recognition problem. Electron Lett. 2010;46(10):682–U632.
59. Sharma A, Paliwal KK. Improved nearest centroid classifier with shrunken distance measure for null LDA method on cancer classification problem. Electron Lett. 2010;46(18):1251–U1225.
60. Sharma A, Paliwal KK. Cancer classification by gradient LDA technique using microarray gene expression data. Data Knowl Eng. 2008;66(2):338–47.
61. Sharma A, Paliwal KK. A gradient linear discriminant analysis for small sample sized problem. Neural Process Lett. 2008;27(1):17–24.
62. Sharma A, Paliwal KK, Onwubolu GC. Class-dependent PCA, MDC and LDA: a combined classifier for pattern classification. Pattern Recogn. 2006;39(7):1215–29.
63. Sharma A, Boroevich K, Shigemizu D, Kamatani Y, Kubo M, Tsunoda T. Hierarchical maximum likelihood clustering approach. In: IEEE transactions on biomedical engineering; 2016. p. 99.
64. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–50.
65. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(2–3):193–218.
66. Vinh NX, Epps J, Bailey J. Information theoretic measures for Clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res. 2010;11:2837–54.
67. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–7.
68. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673–9.
69. Armstrong SA, Staunton JE, Silverman LB, Pieters R. Boer MLd, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet. 2002;30(1):41–7.
70. Yeoh E-J, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell. 2002;1(2):133–43.
71. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001;98(26):15149–54.

72. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. 2002;62(17):4963–7.
73. Kwon OH, Park JL, Kim M, Kim JH, Lee HC, Kim HJ, Noh SM, Song KS, Yoo HS, Paik SG, et al. Aberrant up-regulation of LAMB3 and LAMC2 by promoter demethylation in gastric cancer. Biochem Bioph Res Co. 2011; 406(4):539–45.
74. Du P, Zhang XA, Huang CC, Jafari N, Kibbe WA, Hou LF, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. Bmc Bioinformatics. 2010;11
75. Yamada N, Yasui K, Dohi O, Gen Y, Tomie A, Kitaichi T, Iwai N, Mitsuyoshi H, Sumida Y, Moriguchi M, et al. Genome-wide DNA methylation analysis in hepatocellular carcinoma. Oncol Rep. 2016;35(4):2228–36.
76. Fraley C, Raftery AE. MCLUST version 3 for R: normal mixture modeling and model-based clustering. In: Technical report no 504. USA Seattle, WA: Department of Statistics, University of Washington; 2006. p. 98195–4322.
77. Elhamifar E, Vidal R. Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell. 2013;35(11):2765–81.
78. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705–19.