Supplementary Materials

1.1 Deviation of joint NMF algorithm and its convergence proof

The joint NMF problem minimize the following objective function:

$$\mathcal{F}(W, H_l) = \sum_{l=1}^{3} \| X_l - WH_l \|_F^2.$$ 

with the constraint that $W \geq 0, H_l \geq 0$ and $l = 1, 2, 3$. Similar with the NMF problem, the objective function $\mathcal{F}$ is not convex in both $W$ and $H_l$ together. Therefore, it is unrealistic to design an algorithm to find the global minimum of $\mathcal{F}$. We derive the multiplicative update algorithm which can achieve a local minimum of this problem in the following part.

Based on the simple knowledge of linear algebra, the objective function $\mathcal{F}$ can be reformulated as follows:

$$\mathcal{F}(W, H_l) = \sum_{l=1}^{3} \left[ \text{Tr}(X_l X_l^T) - 2 \text{Tr}(X_l H_l^T W^T) + \text{Tr}(W H_l H_l^T W^T) \right] + \text{Tr}(\Psi W^T) + \sum_{l=1}^{3} \text{Tr}(\Phi_l H_l^T).$$

where $\Psi = [\psi_{ij}]$ and $\Phi_l = [\phi_{ij}^l]$. The partial derivatives of $\mathcal{L}$ with respect to $W$ and $H_l$ are:

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{l=1}^{3} \left[ -2X_l H_l^T + 2WH_l H_l^T \right] + \Psi$$

$$\frac{\partial \mathcal{L}}{\partial H_l} = -2W^T X_l + 2W^T WH_l + \Phi_l, \quad I = 1, 2, 3$$

Based on the KKT conditions $\psi_{ij} W_{ij} = 0$ and $\phi_{ij}^l (H_l)_{ij} = 0$, we get the following equations for $W_{ij}$ and $(H_l)_{ij}$:

$$-\sum_{l=1}^{3} (X_l H_l^T)_{ij} W_{ij} + \sum_{l=1}^{3} (W H_l H_l^T)_{ij} W_{ij} = 0$$

$$-(W^T X_l)_{ij} (H_l)_{ij} + (W^T WH_l)_{ij} (H_l)_{ij} = 0$$

Then we can get the following update rules:

$$W_{ij} \leftarrow W_{ij} \frac{(X_1 H_1^T + X_2 H_2^T + X_3 H_3^T)_{ij}}{(W(H_1 H_1^T + H_2 H_2^T + H_3 H_3^T))_{ij}}$$

$$(H_l)_{ij} \leftarrow (H_l)_{ij} \frac{(W^T X_l)_{ij}}{(W^T W H_l)_{ij}}, \quad I = 1, 2, 3.$$
We have the following theorem to guarantee the convergence of the above update rules and the final solution will be a local optimum.

**Theorem 1** The objective function $F$ of the joint NMF problem is nonincreasing under the above update rules. The objective function is invariant under these updates if and only if $W$ and $H_I$ ($I = 1, 2, 3$) are at a stationary point.

The principle of convergence proof of NMF can be easily used for prove this theorem. Obviously, the objective function $F$ is bounded from below by zero. Here, we just need to show that $F$ is nonincreasing under the update rules. Simply, we just prove the

$$F(ab) = 2(W^T W)_{aa}$$

Because the update is essentially element-wise, it is sufficient to show that each $F_{ab}$ is nonincreasing under the update rule for $H_I$.

**Lemma 1** Function

$$G(h, (H_I)_{ab}^{(t)}) = F_{ab}((H_I)_{ab}^{(t)}) + F_{ab}'((H_I)_{ab}^{(t)}) (h - (H_I)_{ab}^{(t)}) + \frac{(W^T WH_I)_{ab}}{(H_I)_{ab}^{(t)}} (h - (H_I)_{ab}^{(t)})^2$$

is an auxiliary function for $F_{ab}$.

**Proof** Obviously, $G(h, h) = F_{ab}(h)$. Here we only show that $G(h, (H_I)_{ab}^{(t)}) \geq F_{ab}(h)$. To achieve this, we compare the Taylor series expansion of $F_{ab}(h)$

$$F_{ab}(h) = F_{ab}((H_I)_{ab}^{(t)}) + F_{ab}'((H_I)_{ab}^{(t)}) (h - (H_I)_{ab}^{(t)}) + (W^T W)_{aa} (h - (H_I)_{ab}^{(t)})^2$$

with the auxiliary function $G(h, (H_I)_{ab}^{(t)})$ to find that $G(h, (H_I)_{ab}^{(t)}) \geq F_{ab}(h)$ is equivalent to

$$\frac{(W^T WH_I)_{ab}}{(H_I)_{ab}^{(t)}} \geq (W^T W)_{aa}.$$ 

Obviously, we have

$$(W^T WH_I)_{aa} = \sum_k (W^T W)_{ak} (H_I)_{kb}^{(t)} \geq (H_I)_{ab}^{(t)} (W^T W)_{aa}.$$ 

Thus $G(h, (H_I)_{ab}^{(t)}) \geq F_{ab}(h)$ holds.
Proof of Theorem 1 We can get the following update rule based on the auxiliary function $G(h, (H_I)_t)$:

$$(H_I)_t^{(t+1)} = (H_I)_t^{(t)} - \frac{(-2W^TX_I + 2W^TH_I)_ab}{2(W^TH_I)_ab} = (H_I)_t^{(t)} \frac{(W^TX_I)_ab}{(W^TH_I)_ab}, \quad I = 1, 2, 3.$$  

Due to the property of the auxiliary function $G(h, (H_I)_t)$ for $F_{ab}$, $F_{ab}$ is nonincreasing under this update rule.

1.2 Simulation

A multi-dimensional data set which consists of a ‘miRNA expression’ matrix $X_1$, a ‘miRNA expression’ matrix $X_2$ and a ‘gene-expression’ matrix $X_3$ was generated such that they contain both independent modules and md-modules (Figure 2). We assume that the underlying patterns among DNA methylation markers, miRNA and gene expression profiles show coherent characteristic and which can be reduced to component vectors among $W$ (feature matrix). These component vectors form the basis vectors and their corresponding associated vectors in $H_1$, $H_2$ and $H_3$ (coefficient matrices) form the md-modules’ membership correspondingly. We use three matrices to characterize the properties of all these factors. Note that for simplicity we only describe the component and association vectors in a binary fashion. The in-silico DNA methylation, microRNA and gene expression matrices are then computed as follows: $WH_1 = X_1$, $WH_2 = X_2$ and $WH_3 = X_3$. Finally, Gaussian noise is added to the matrices. We used these models to simulate data sets in order to investigate the proposed to handle multiple data sets. In our simulations, the expression data are simulated for different variables within 45 samples. We allowed for a total of 4 component vectors correspond to four md-modules.

A simulated data set with the same number of samples (rows) and different number of features (columns) was generated. The joint NMF method can accurately discover the patterns embedded in these data. A pattern may involve as many as all 3 datasets simultaneously or only cover two datasets. These different patterns may share the same samples (overlap) or/and the same features.

We computed the precision rate of the identified md-modules with that embedded in the simulated matrices. We observed that our method can well identify the underlying patterns even with relative high noise level (Figure S1).

1.3 Two key parameters $K$ and $T$

The lower dimensionality $K$ of the reduced space is a key parameter for this study. The choice of $K$ is often problem-dependent and is a long-standing open problem. Although it has been discussed in many dimensionality-reduction problems and/or clustering problems, it is still a challenging issue now. We should note that the approach employing consensus clustering and a cophenetic coefficient to measure stability and further to select the optimal $K$ as done in Brunet et al. 2004 is inappropriate to our study. Since the goal of the consensus clustering and a cophenetic coefficient are designed for the traditional clustering task, while our goal is try to extract subtle local patterns among (multiple) matrices. An essential feature of the joint NMF approach is that it reduces the dataset to a lower common dimensional NMF space. Initial calculations and enrichment analysis were performed to select an appropriate size $K$. Three empirical factors guided the selection of the dimensionality $K$: 1) The changing trend of errors between the input matrices and our model reconstructed data with different dimensionality; 2) The significant rate of permutation tests for statistical test for vertical correlations; 3) The performance of enrichment analysis compared with
Figure S1: The proposed method was tested on the simulated data sets to see if it can identify the embedded multi-dimensional modules. These data were generated according to a simple model: \( WH_1 = X_1, WH_2 = X_2 \) and \( WH_3 = X_3 \) and Gaussian noise is added to the matrices with deviation controlling the noise level. Performance was evaluated in terms of the precision ratio when comparing the members of the identified patterns with the real ones. The figure shows the performance as a function of the noise level. We can see that the proposed model can well identify the modular patterns even with noise level \( = 1 \).

random generated modules. Finally, we set the reduced dimension of the matrix factorization \( k \) to 200, approximately equal to the number of KEGG pathways in our enrichment analysis.

To select the proper threshold \( T \), we assess the enrichment rate of gene modules with respect to GO biological process. For comparison, the mean rate of functional enrichment for 100 corresponding random runs was also calculated. In Figure S2, we show the fold change (the ratio of real enrichment rate against the mean of 100 random ones) for different cutoff \( T \). The highest peak shows that the \( T = 5 \) is a proper threshold.
Figure S2: The distribution of the enrichment ratio of gene modules and the mean of that for modules of 100 random runs under different cutoff $T$ was shown in the top subfigure, and the corresponding fold change was shown in the below subfigure.
1.4 Size distribution of md-modules

Figure S3 illustrates the module size distributions of md-modules in each dimensions.

![Size distributions of md-modules](image)

Figure S3: Size distributions of md-modules with threshold $T = 5$.

1.5 Package

We have implemented the method as a Matlab software package, which is available in ‘MDmodule.toolbox.zip’.

1.6 Dataset

The dataset used in this paper is available in ‘TCGA.Data.zip’.