# Consistent Multiple Nonnegative Matrix Factorization with Hierarchical Information for Gene Functional Modules Mining

## Supporting Information

Files in this Data Supplement:

- SI Parameter Tuning
- SI Measurement Definition
- SI $\mathcal{M}_{sim}$
- SI Significant Analysis
- SI AUC
- SI Supervised CMNMF

### SI Parameter Tuning

we use the Euclidean distance as the gene similarity measure to cluster the genes; for Kernel-Kmeans, Gaussian Kernel function is used for calculating the gene similarity, there is a hyper-parameter $\sigma$ in Kernel-Kmeans, we search it in {0.001,0.01,0.1,1,10}; we set the values of hyper-parameters $\alpha$ and $\beta$ in L-DA as [1] suggested. Because there are common sparse constraints in NMF, HMF , ColNMF, and CMNMF, we search the parameters $\lambda_1$, $\lambda_2$ in a grid {0.001,0.01.0.1,0,1,10}.

### SI Measurement Definition

In Our paper, we evaluate the gene clustering results from three external criteria. These criteria includes $F_1$, $Jaccard\ Index$ and $Rand\ Index$.

We use $n$ to denote the total gene number, all genes can be denoted as $S = \{O_1, ..., O_n\}$, we use gene pathways $X = \{X_1, ..., X_r\}$ as the ground-truth partition of genes, and use clustering result $Y = \{Y_1, ..., Y_s\}$ as prediction partitions, then define the following notations:

- TP, the number of pairs of elements in $S$ that are in the same set in $X$ and in the same set in $Y$
- TN, the number of pairs of elements in $S$ that are in different sets in $X$ and in different sets in $Y$
- FN, the number of pairs of elements in $S$ that are in the same set in $X$ and in different sets in $Y$
- FP, the number of pairs of elements in $S$ that are in different sets in $X$ and in the same set in $Y$

Now we can define the evaluation measures as below:

$$F_1\ measure = \frac{2PR}{P+R} \quad (P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN})$$
$$Jaccard\ Index = \frac{TP}{TP+FP+FN}$$
$$Rand\ Index = \frac{TP+TN}{TP+FP+FN+TN}$$

## SI $\mathcal{M}_{sim}$

We use $\mathcal{M}_{sim}$ [2] as an internal criteria to evaluate gene clustering results. we take $\boldsymbol{E}[(Sim)_{inter}]$ to denote the average similarity between different gene clusters, $\boldsymbol{E}[(Sim)_{intra}]$ to denote average similarity of different genes within gene clusters. then $\mathcal{M}_{sim}$ is defined as:

$$\mathcal{M}_{sim} = \frac{\boldsymbol{E}[(Sim)_{intra}]}{\boldsymbol{E}[(Sim)_{inter}]}$$

where

$$\boldsymbol{E}[(Sim)_{intra}] = \frac{1}{|Y|}\sum_{i=1}^{|Y|}\frac{\sum_{j=1}^{|Y_i|}\sum_{k=j+1}^{|Y_i|} FS_{simGIC}(G_j, G_k)}{\binom{|Y_i|}{2}}$$
$$\boldsymbol{E}[(Sim)_{inter}] = \frac{1}{\binom{|Y|}{2}}\sum_{\forall(i,j)\in P}\frac{\sum_{k=1}^{|Y_i|}\sum_{l=1}^{|Y_j|} FS_{simGIC}(G_k, G_l)}{|Y_i||Y_j|}$$

, $P = \{(i,j)|1 \le i < j \le |Y|\}$, $FS_{simGIC}(G_k, G_l)$ [3] is used to measure the similarity between gene $G_k$ and $G_l$ by GO terms.

The $\mathcal{M}_{sim}$, an internal clustering evaluation measure, is used for evaluating the performance of clustering genes by factorizing gene-phenotype associations. The higher the $\mathcal{M}_{sim}$, the better the performance. For $\mathcal{M}_{sim}$, genes are featured with annotated GO terms, and $FS_{simGIC}(G_k, G_l)$ [3] is used to measure the similarity between gene $G_k$ and $G_l$ by GO terms. In other words, genes are clustered well if the GO annotations of genes in a cluster are similar.

## SI Significant Analysis

Table 1: The Student's t-test results of each pair methods on $F_1$ measure

| | HAC | Kmeans | KK | LDA | NMF | HMF | ColNMF | CMNMF |
|---|---|---|---|---|---|---|---|---|
| HAC | - | 5.90E-09 | 8.82E-10 | 0 | 1.52E-10 | 6.85E-05 | 8.39E-08 | 5.94E-09 |
| Kmeans | 5.90E-09 | - | 8.40E-03 | 1.75E-14 | 4.97E-12 | 1.19E-09 | 8.36E-11 | 1.11E-10 |
| KK | 8.82E-10 | 8.40E-03 | - | 5.67E-14 | 2.95E-12 | 2.15E-10 | 2.61E-11 | 2.99E-11 |
| LDA | 0 | 1.75E-14 | 5.67E-14 | - | 2.29E-03 | 5.00E-10 | 6.59E-06 | 1.25E-05 |
| NMF | 1.52E-10 | 4.97E-12 | 2.95E-12 | 2.29E-03 | - | 1.86E-05 | 3.58E-02 | 3.02E-06 |
| HMF | 6.85E-05 | 1.19E-09 | 2.15E-10 | 5.00E-10 | 1.86E-05 | - | 5.70E-03 | 7.25E-08 |
| ColNMF | 8.39E-08 | 8.36E-11 | 2.61E-11 | 6.59E-06 | 3.58E-02 | 5.70E-03 | - | 6.53E-07 |
| CMNMF | 5.94E-09 | 1.11E-10 | 2.99E-11 | 1.25E-05 | 3.02E-06 | 7.25E-08 | 6.53E-07 | - |

Table 2: The Student's t-test results of each pair methods on *Jaccard Index* measure

| | HAC | Kmeans | KK | LDA | NMF | HMF | ColNMF | CMNMF |
|---|---|---|---|---|---|---|---|---|
| HAC | - | 3.93E-02 | 5.38E-05 | 0 | 7.65E-16 | 8.66E-12 | 5.59E-15 | 6.26E-11 |
| Kmeans | 3.93E-02 | - | 2.65E-05 | 2.10E-16 | 2.57E-12 | 1.08E-08 | 3.17E-11 | 2.37E-10 |
| KK | 5.38E-05 | 2.65E-05 | - | 8.04E-16 | 2.84E-13 | 7.76E-11 | 1.64E-12 | 1.99E-11 |
| LDA | 0 | 2.10E-16 | 8.04E-16 | - | 1.05E-05 | 2.80E-10 | 7.59E-09 | 4.90E-05 |
| NMF | 7.65E-16 | 2.57E-12 | 2.84E-13 | 1.05E-05 | - | 3.05E-05 | 1.58E-02 | 3.16E-06 |
| HMF | 8.66E-12 | 1.08E-08 | 7.76E-11 | 2.80E-10 | 3.05E-05 | - | 5.40E-03 | 7.91E-08 |
| ColNMF | 5.59E-15 | 3.17E-11 | 1.64E-12 | 7.59E-09 | 1.58E-02 | 5.40E-03 | - | 5.31E-07 |
| CMNMF | 6.26E-11 | 2.37E-10 | 1.99E-11 | 4.90E-05 | 3.16E-06 | 7.91E-08 | 5.31E-07 | - |

We give a significant analysis of the results of each pair methods, Table 1, Table 2, Table 3 give the Student's t-test on $F_1$ measure, *Jaccard Index* and $\mathcal{M}_{sim}$ respectively. The last columns of each table show that the result of our model CMNMF outperforms other baselines significantly ($p<0.01$).

Table 3: The Student's t-test results of each pair methods on $\mathcal{M}_{sim}$

| | HAC | Kmeans | KK | LDA | NMF | HMF | ColNMF | CMNMF |
|---|---|---|---|---|---|---|---|---|
| HAC | - | 3.42E-05 | 7.06E-07 | 0 | 6.11E-11 | 1.80E-09 | 9.17E-09 | 5.33E-04 |
| Kmeans | 3.42E-05 | - | 1.89E-03 | 4.15E-09 | 1.66E-02 | 2.79E-07 | 1.65E-01 | 2.38E-06 |
| KK | 7.06E-07 | 1.89E-03 | - | 6.76E-14 | 3.81E-02 | 3.57E-04 | 9.29E-03 | 1.42E-07 |
| LDA | 0 | 4.15E-09 | 6.76E-14 | - | 7.08E-10 | 7.93E-04 | 5.07E-10 | 3.90E-18 |
| NMF | 6.11E-11 | 1.66E-02 | 3.81E-02 | 7.08E-10 | - | 1.49E-06 | 1.83E-01 | 9.94E-11 |
| HMF | 1.80E-09 | 2.79E-07 | 3.57E-04 | 7.93E-04 | 1.49E-06 | - | 6.08E-07 | 7.25E-10 |
| ColNMF | 9.17E-09 | 1.65E-01 | 9.29E-03 | 5.07E-10 | 1.83E-01 | 6.08E-07 | - | 3.68E-09 |
| CMNMF | 5.33E-04 | 2.38E-06 | 1.42E-07 | 3.90E-18 | 9.94E-11 | 7.25E-10 | 3.68E-09 | - |

## SI AUC

AUC, area under the curve (AUC) is the area under the ROC (Receiver Operating Characteristic) curve. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. In our paper, we use AUC-K (like AUC50) to calculate the area under the ROC from the zero point, which counts only total K false positive samples. AUCALL means we take all false positive samples into account, which equals to the AUC score we usually refer to.

## SI Supervised CMNMF

By considering prior gene classification prior, our proposed CMNMF can be utilized for another task, multi-label gene signal pathway classification problem as well. We call it Supervised CMNMF (S-CMNMF) as including gene pathway classification prior. Thus, the loss function of S-CMNMF can be rewrite as:

$$\begin{aligned} L_s = & \left\|\boldsymbol{A}_1 - \boldsymbol{G}\boldsymbol{P}_1\right\|_F^2 + \alpha \left\|\boldsymbol{A}_2 - \boldsymbol{G}\boldsymbol{P}_2\right\|_F^2 - \beta tr(\boldsymbol{P}_1\boldsymbol{M}\boldsymbol{P}_2^T) \\ & + \gamma \left\|\boldsymbol{G} - \boldsymbol{G}_0\right\|_F^2 \\ & \sum_j \boldsymbol{G}_{ij} = 1, \quad \sum_i (\boldsymbol{P}_1)_{ij} = 1, \quad \sum_i (\boldsymbol{P}_2)_{ij} = 1 \end{aligned} \tag{1}$$

### Computation of $G$ in Supervised CMNMF

We fix variables $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, the partial derivative of Equation (1) with respect to $\boldsymbol{G}$ is:

$$\frac{\partial L_s}{\partial \boldsymbol{G}} = -2(\boldsymbol{A}_1\boldsymbol{P}_1^T - \boldsymbol{G}\boldsymbol{P}_1\boldsymbol{P}_1^T) - 2\alpha(\boldsymbol{A}_2\boldsymbol{P}_2^T - \boldsymbol{G}\boldsymbol{P}_2\boldsymbol{P}_2^T) + 2\gamma(\boldsymbol{G} - \boldsymbol{G}_0)$$

the multiplicative update rule is:

$$\boldsymbol{G}_{ij} \leftarrow \boldsymbol{G}_{ij} \frac{(\boldsymbol{A}_1 \boldsymbol{P}_1^T + \alpha \boldsymbol{A}_2 \boldsymbol{P}_2^T + \gamma \boldsymbol{G}_0)_{ij}}{(\boldsymbol{G}\boldsymbol{P}_1 \boldsymbol{P}_1^T + \alpha \boldsymbol{G}\boldsymbol{P}_2 \boldsymbol{P}_2^T + \gamma \boldsymbol{G})_{ij}}$$

To satisfy the equality constraint, we normalize $\boldsymbol{G}_{ij}$ as $\boldsymbol{G}_{ij} \leftarrow \frac{\boldsymbol{G}_{ij}}{\sum_j \boldsymbol{G}_{ij}}$..

**Computation of $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ in Supervised CMNMF**

Because the updating rules of supervised CMNMF for $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are the same as CMNMF above, we would not present it here again.

**The Algorithm of Supervised CMNMF**

For supervised CMNMF algorithm, we just need to change the update rule for $\boldsymbol{G}$ as we talked in Computation of $\boldsymbol{G}$ in Supervised CMNMF part, the rest is the same as CMNMF.

# References


[1] Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06, p. 178. ACM Press, New York, New York, USA (2006). http://dl.acm.org/citation.cfm?id=1148170.1148204

[2] Bordino, I., Castillo, C., Donato, D., Gionis, A.: Query similarity by projecting the query-flow graph. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10, p. 515. ACM Press, New York, New York, USA (2010). doi:10.1145/1835449.1835536. http://dl.acm.org/citation.cfm?id=1835449.1835536

[3] Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C., Xuan, P.: Measuring gene functional similarity based on group-wise comparison of GO terms. Bioinformatics (Oxford, England) **29**(11), 1424–32 (2013). doi:10.1093/bioinformatics/btt160