Non-negative Matrix Factorization

Dongfang Wang

November 29, 2013

Outline

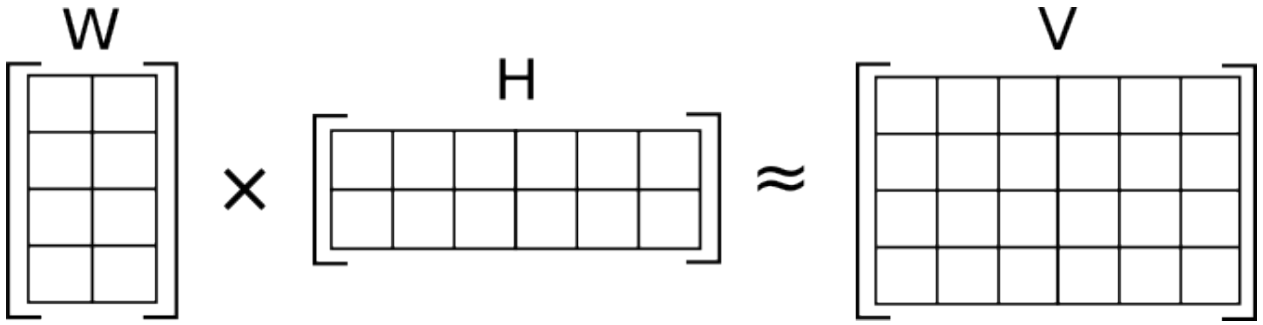
Outline

Definition

Non-negative matrix factorization:

$$\mathbf{V} \approx \mathbf{WH}$$

1. $\mathbf{V}$ is an $n \times m$ matrix, and $\mathbf{M} \in R^{n \times r}$, $\mathbf{H} \in R^{r \times m}$.



2. all elements of the three matrices are **non-negative**
3. $\mathbf{W}$: every column is a **basis image**, **building block**
   $\mathbf{H}$: **encoding**; coefficients of the linear combination of the building block.

$$\vec{V}_i = \mathbf{W}\vec{H}_i$$

Learn the parts of object

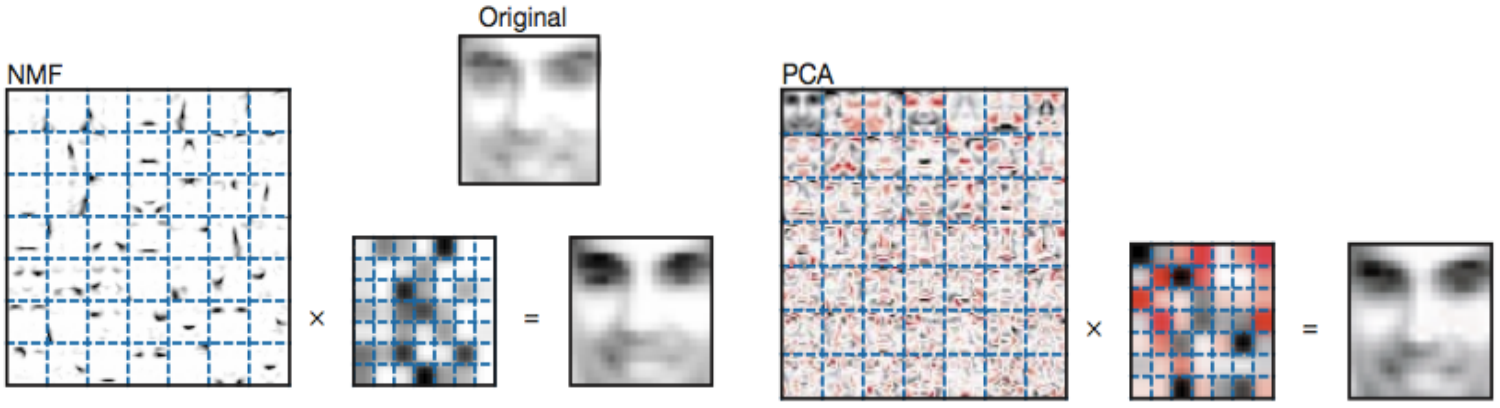


Figure : Lee and Seung(1999)

Different Constraints

- PCA : orthogonal columns of $W, H \Longrightarrow$ **"eigenfaces"**
- NMF : non-negative elements W, H
  $\Longrightarrow$ **only additive combinations**
  $\Longrightarrow$ **combining parts to form a whole**

Algorithm

Cost Functions:

- "distance":

$$||A-B||_2^2 = \sum_{ij}(A_{ij}-B_{ij})^2$$

- "divergence":

$$D(A||B) = \sum_{ij}(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij})$$

Multiple Update Rules

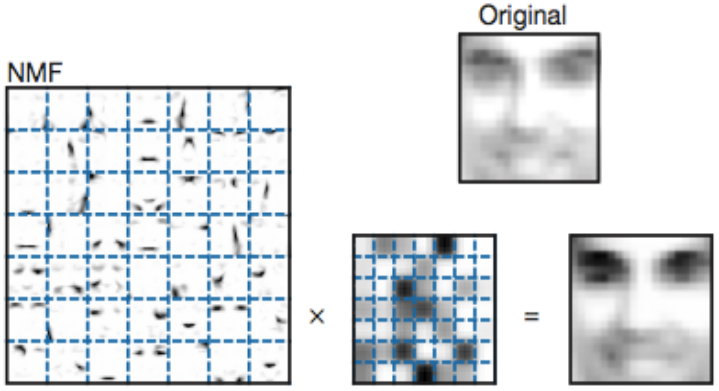
$$H_{ij} \leftarrow H_{ij}\frac{(W^TV)_{ij}}{(W^TWH)_{ij}}$$

$$W_{ij} \leftarrow W_{ij}\frac{(VH^T)_{ij}}{(WHH^T)_{ij}}$$

NMF and Sparsity

NMF $\Rightarrow$ sparsity

- $\mathbf{W}$ is sparse : non-global and contain serveral versions of parts.
- $\mathbf{H}$ is sparse : any given example dosen't use all available parts.



Need of explicit sparseness contraints

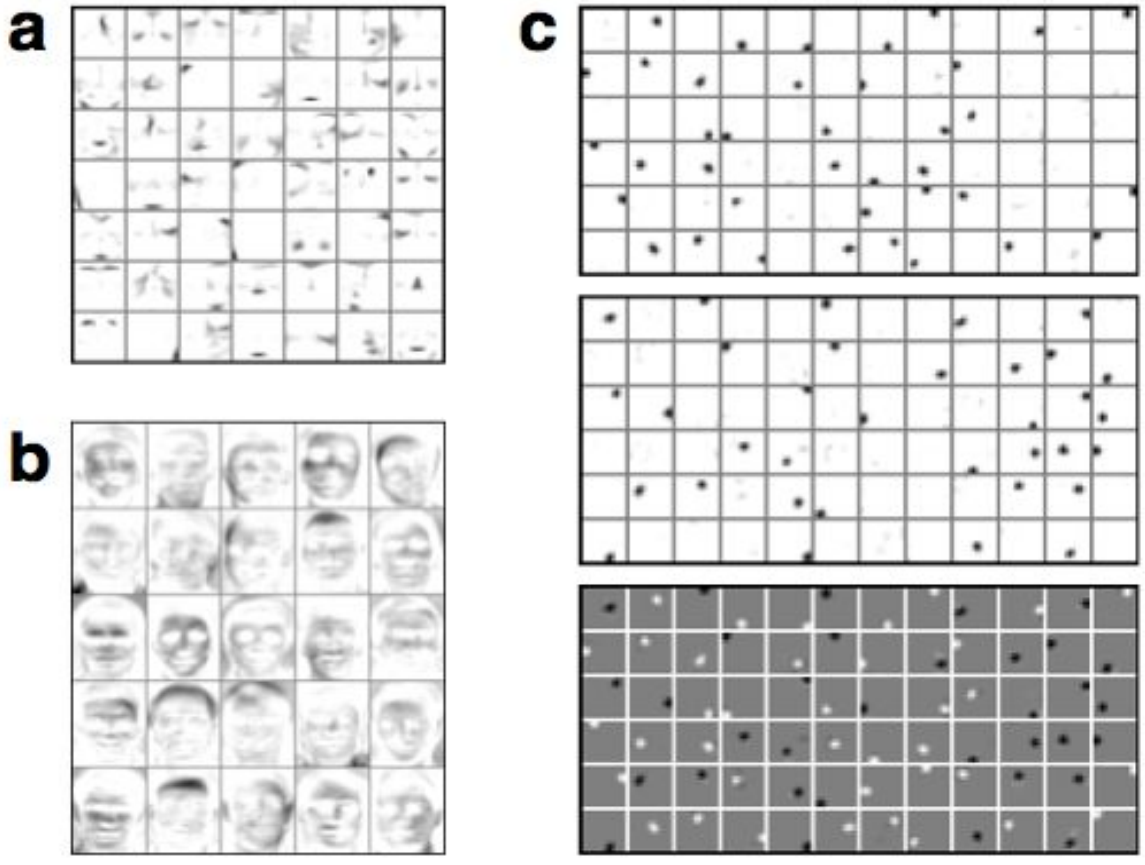


Figure : Hoyer (2004)

What exactly should be sparse ?

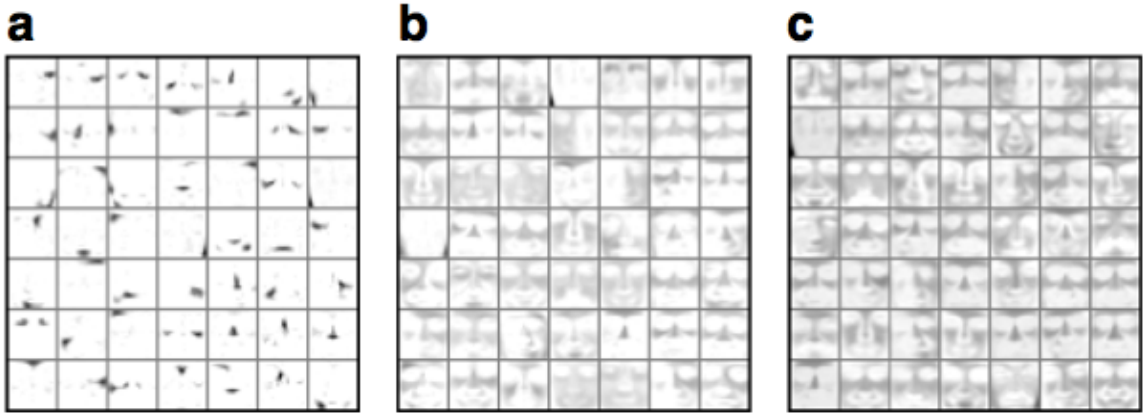


Figure : Hoyer (2004)

a large sparseness of $\mathbf{W}$ and no constraint of $\mathbf{H}$

b small sparseness of $\mathbf{W}$ and no constraint of $\mathbf{H}$

c large sparseness of $\mathbf{H}$ and no constraint of $\mathbf{W}$

This is a question that cannot be given a general answer.

1. a doctor analyzing disease patterns might assume that most diseases are rare (hence sparse) but that each disease can cause a large number of symptoms. Assuming that symptoms make up the rows of her matrix and the columns denote different individuals, in this case it is the coefficients which should be sparse and the basis vectors unconstrained.

Example

$\mathbf{V}$: symptoms $\times$ individuals
$\mathbf{W}$: symtoms $\times$ diseases
$\mathbf{H}$: coffeients $\times$ individuals

2. when trying to learn useful features from a database of images, it might make sense to require both W and H to be sparse, signifying that any given object is present in few images and affects only a small part of the image.

Outline

Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data

Philip M. Kim and Bruce Tidor

Genome Res. 2003 13: 1706-1718
Access the most recent version at doi:10.1101/gr.903503

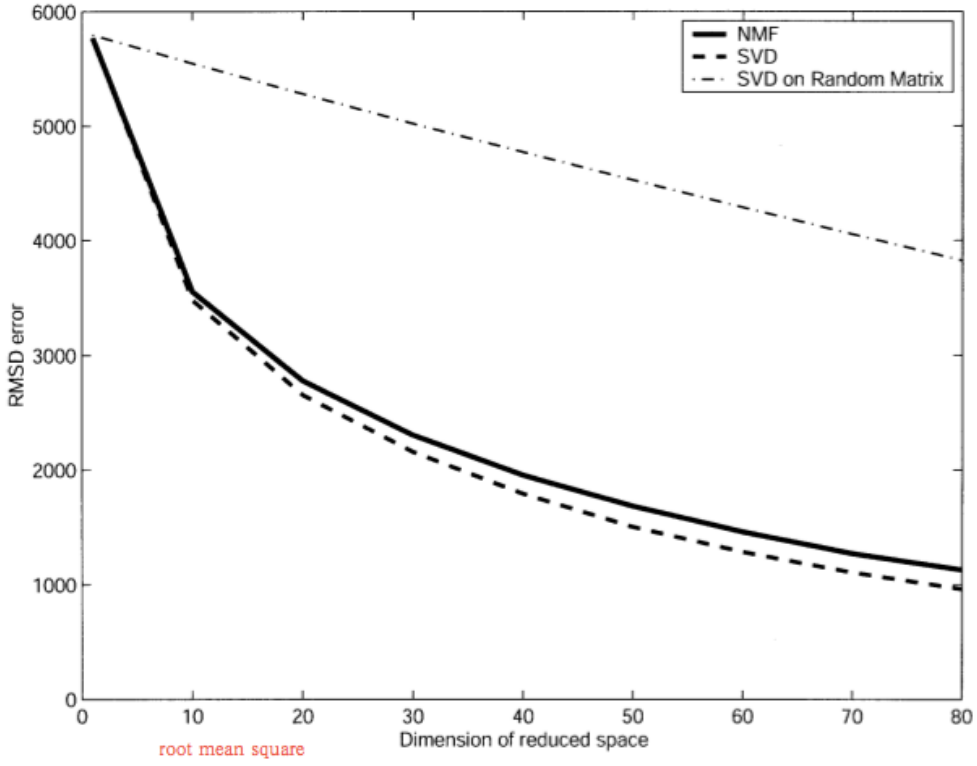
Data preprocessing

Log-transformed ratios can be positive or negative

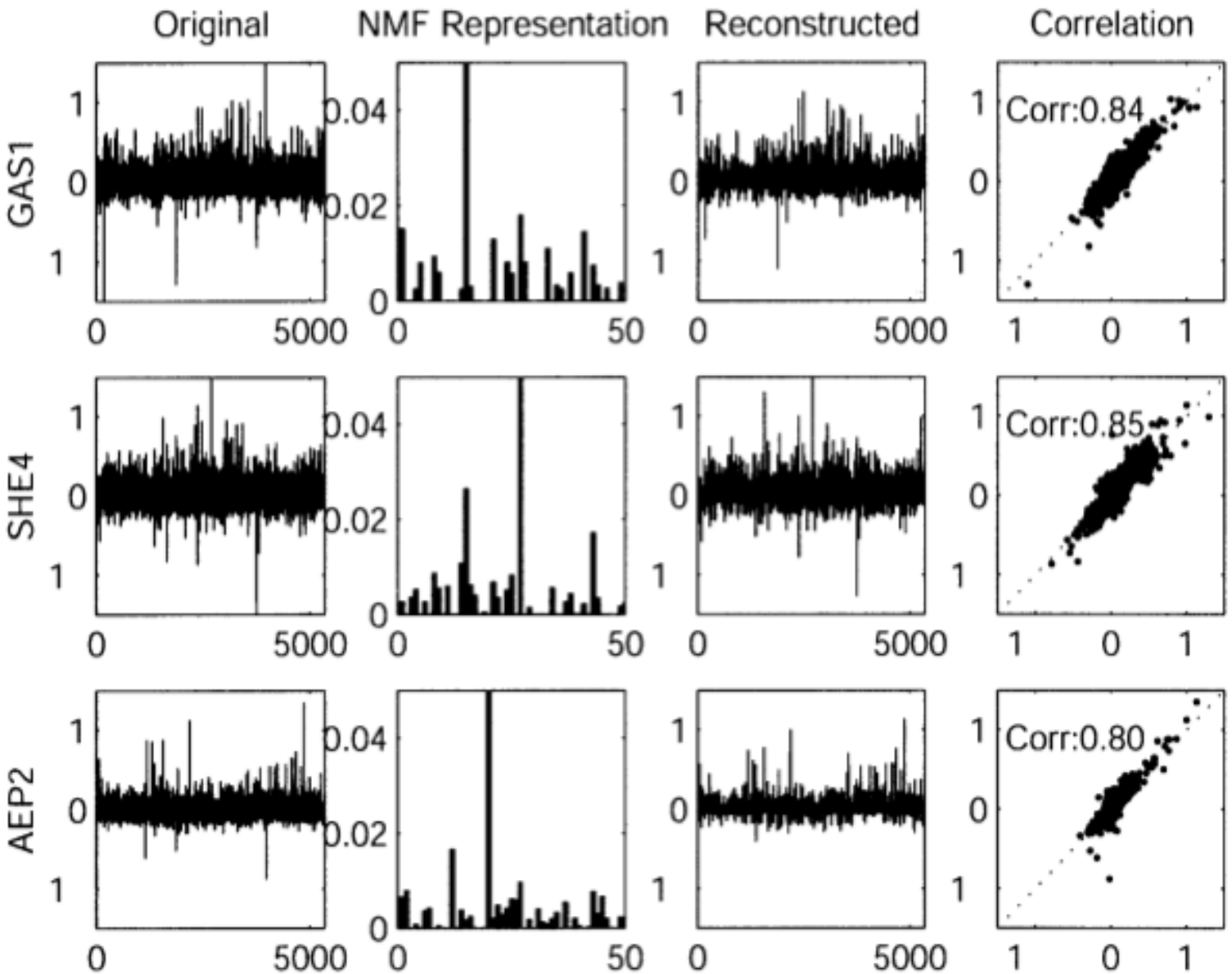
1. Fold the data. Each gene is represented in two rows (or columns)
   - positive: value + 0
   - negative: 0 + absolute value
2. NMF performs most optimally on sparse data sets

Selection of NMF Dimensionality

$$r < \frac{mn}{m+n}$$



1. **SVD:** produce the minimum error for a given dimensionality
2. **Random matrix:** assumed to be unstructured data
3. **Selection:** not much steeper than the unstructured data ?


Original
NMF Representation
Reconstructed
Correlation
GAS1
SHE4
AEP2
Corr:0.84
Corr:0.85
Corr:0.80
0
5000
50
0.04
0.02
1
0

Robust

Different Starting points

- The coorelation coefficient is found to be > 0.9 between different starting points pairs.

Robustness to Noise

Noise added	NMF basis vectors	Reconstructed data	Original data
0.2	0.933	0.930	0.943
0.5	0.879	0.893	0.781
1	0.865	0.816	0.573
5	0.368	0.313	0.159

Outline

Bioinfomatics-2011

Aim

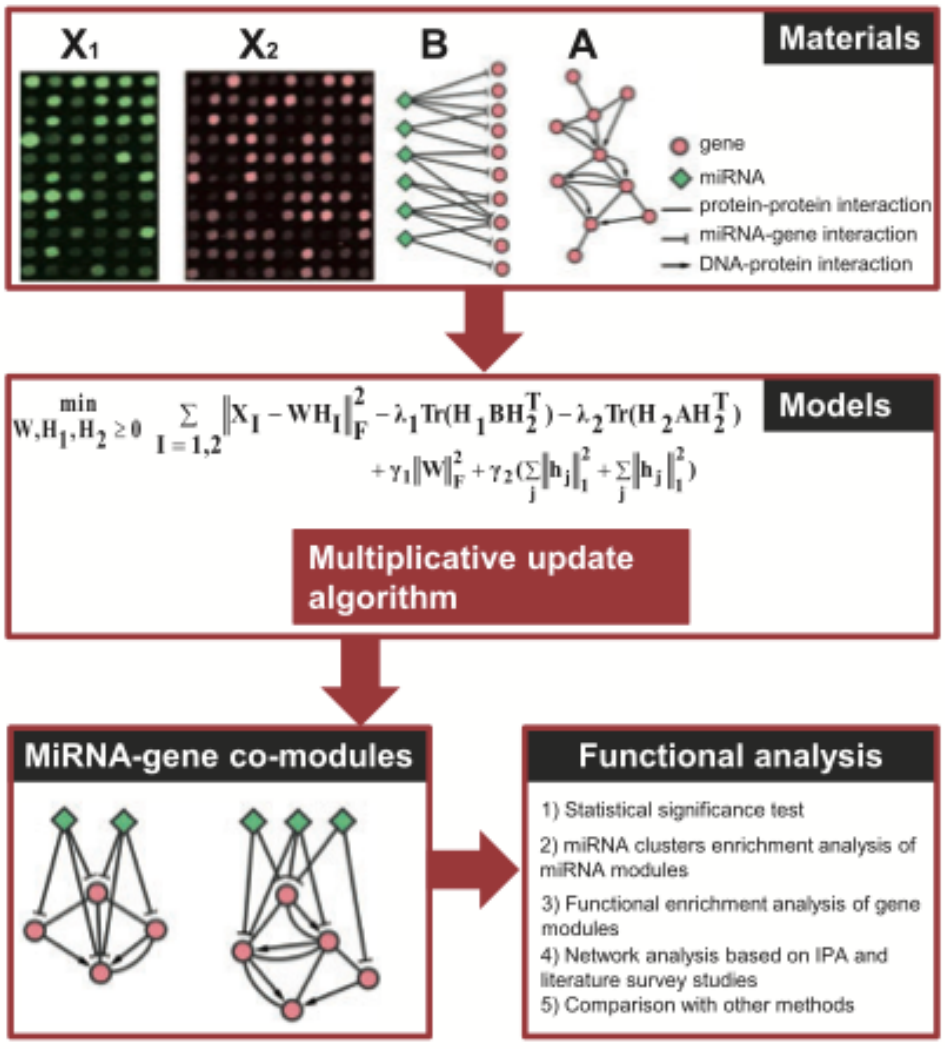
reconstruct miRNA regulatory modules based on the integration of multiple genomic data sources

One gene $\leftarrow$ multiple miRNAs
One miRNA $\rightarrow$ multiple genes
$\Rightarrow$ **a miRNA-gene comodule:** a set of one miRNAs and their co-regulated genes

Data Source

1. predicted miRNA-gene interactions
2. the expression profiles of miRNAs and genes
3. gene-gene interaction network constructed by PPI and DNA-protein interaction network

Computational framework



Multiple NMF

$$\min_{W,H_1,H_2\geq 0} \quad \sum_{i=1,2} ||X_i - \mathbf{W}H_i||_F^2$$
$$-\lambda_1 tr(H_2 A H_2^T) - \lambda_2 tr(H_1 B H_2^T)$$
$$+\gamma_1 ||W||_F^2 + \gamma_2 (\sum_j ||h_j^1||_1^2 + \sum_j ||h_j^2||_1^2)$$

Objective function of modeling miRNA and gene expression profiles

$$\min_{W,H>0} \quad ||X - WH||_F^2$$

$$\min_{W,H_1,H_2>0} \quad \sum_{i=1,2} ||X_i - \mathbf{W}H_i||_F^2$$

common basis matrix: coordinated miRNA-gene comodules
different coeffient matrices

Multiple NMF

$$\min_{W,H_1,H_2 \geq 0} \quad \sum_{i=1,2} ||X_i - \mathbf{W}H_i||_F^2$$
$$-\lambda_1 tr(H_2 A H_2^T) - \lambda_2 tr(H_1 B H_2^T)$$
$$+\gamma_1 ||W||_F^2 + \gamma_2 (\sum_j ||h_j^1||_1^2 + \sum_j ||h_j^2||_1^2)$$

Network-regularized constraints

1. A : gene interaction network

$$\max \sum_{ij} a_{ij} (h_i^2)^T h_j^2 = tr(H_2 A H_2^T)$$

2. B : a bipartite miRNA-gene network

$$\max \sum_{ij} b_{ij} (h_i^1)^T h_j^2 = tr(H_1 B H_2^T)$$

Multiple NMF

$$\min_{W,H_1,H_2\geq 0} \quad \sum_{i=1,2} ||X_i - \mathbf{W}H_i||_F^2$$
$$-\lambda_1 tr(H_2 A H_2^T) - \lambda_2 tr(H_1 B H_2^T)$$
$$+\gamma_1 ||W||_F^2 + \gamma_2 (\sum_j ||h_j^1||_1^2 + \sum_j ||h_j^2||_1^2)$$

Sparse NMNMF

1. limit the growth of W : $||W||_F^2$
2. make the coefficient matrices sparse: $\sum_j ||h_j^1||_1^2 + \sum_j ||h_j^2||_1^2$

MiRNA-gene Comodule Assignment

Some genes may be active in multiple modules and others may not participate in any module

$z-$score : each element based on the rows of H_1 and H_2:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i}$$

Biological Significance of the Comodule

1. The anti-correlations between miRNAs and genes within a comodule are statistically significant in 69.4% of the modules
2. Enriched in genomic miRNA clusters and known functional sets

Table 1. Summary of miRNA modules that are enriched in miRNA clusters

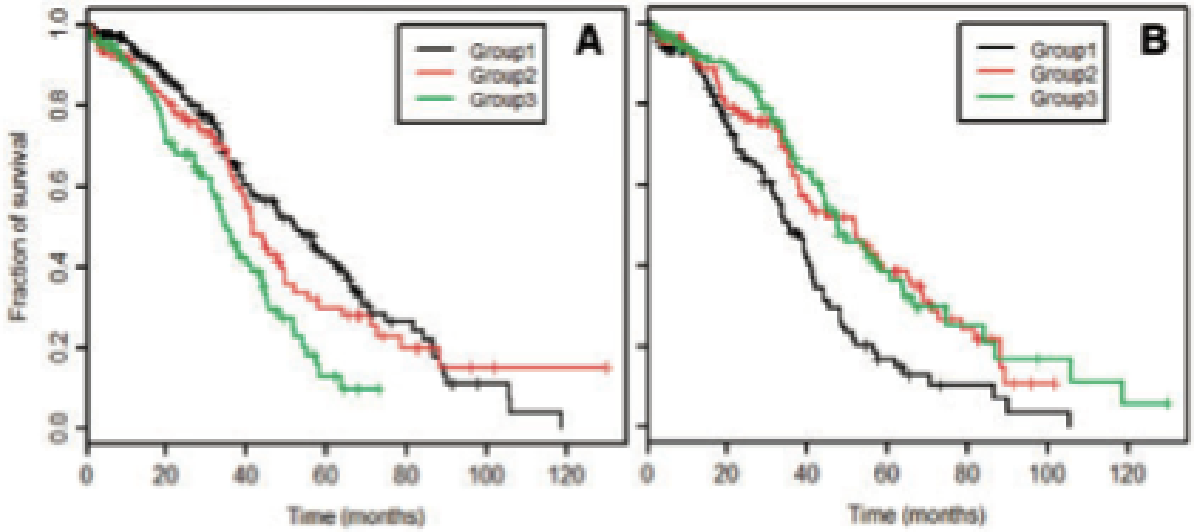
No.	*q*-value	Overlap miRNAs	Loci	FS
10	0.002	mir-449b, mir-449a	5q11.2	Yes
	0.001	mir-34b*, mir-34c-5p	11q23.1	Yes
14	0.002	mir-143, mir-145	5q32	Yes
16	3.94e-05	mir-182*, mir-96, mir-183	7q32.2	Yes
17	0.001	mir-144, mir-451	17q11.2	Yes
18	0.001	mir-452, mir-224	Xq28	No
19	0.005	mir-30b*, mir-30d*, mir-30d, mir-30b	8q24.22	Yes
20	1.97e-5	mir-96, mir-183, mir-182	7q32.2	Yes
42	0.005	mir-199a-5p, mir-214	1q24.3	Yes
46	0.001	mir-144, mir-451, mir-144*	17q11.2	Yes
48	6.78e-12	mir-513b, mir-513c, mir-508-3p, mir-506, mir-507, mir-509-3-5p, mir-514, mir-509-3p, mir-509-5p	Xq27.3	No
50	0.008	mir-502-3p, mir-500*	Xp11.23	No

Table 2. Functional analysis of selected miRNA-gene comodules

No.	GO biological process terms	CG	PT	Cancer miRNAs	Num	OC miRNAs
7	Immune system process; regulation of cell activation; regulation of cell proliferation	Yes	4.4e-165	mir-142-5p, mir-142-3p, mir-21*	3/3	mir-21*
15	Immune response; immune system process; defense response; inflammatory response; response to external stimulus; cell activation	Yes	8.6e-254	mir-142-5p, mir-142-3p, mir-150, mir-146a	4/4	
23	Negative regulation of immune system; response to external stimulus; regulation of cell division; cell adhesion; regulation of cell migration; cell communication;	Yes	1.9e-151	mir-22, mir-199a-5p, mir-145, mir-10b	4/5	mir-22, mir-199a-5p, mir-145, mir-10b
25	Calcium-dependent cell–cell adhesion; synaptic transmission; cell adhesion; extracellular structure organization		4.2e-4	mir-10b*, mir-135b, mir-10b	3/4	mir-10b*, mir-10b

Analysis

1. Biological significance of the comodules
2. Clinical Characterization Based on the Basis Matrix
   One parent $\Leftrightarrow$ One row in W



Published online 8 August 2012

Nucleic Acids Research, 2012, Vol. 40, No. 19 ***9379–9391***
doi:10.1093/nar/gks725

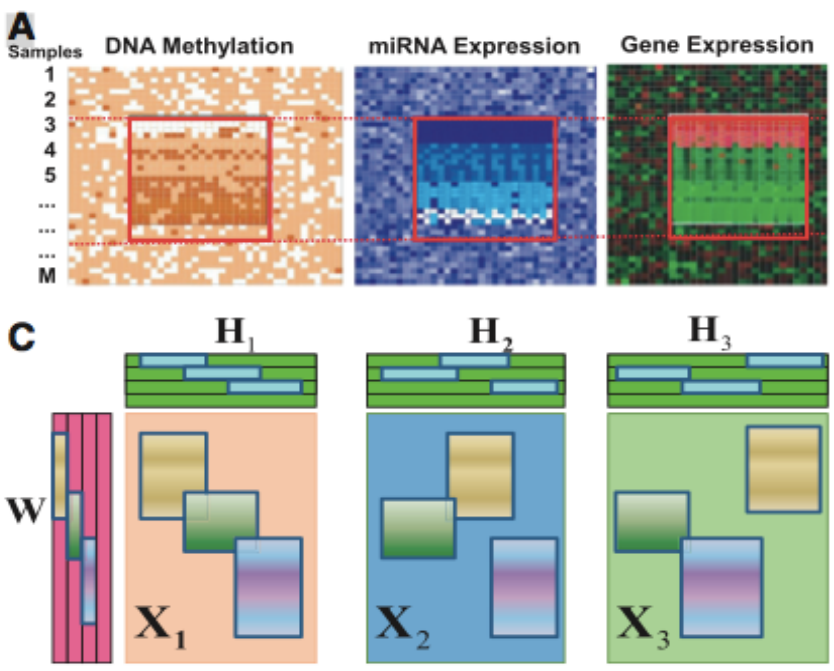
Discovery of multi-dimensional modules by integrative analysis of cancer genomic data

Shihua Zhang[1,2], Chun-Chi Liu[3], Wenyuan Li[1], Hui Shen[4], Peter W. Laird[4] and Xianghong Jasmine Zhou[1,*]

Joint Factorization Framework

1. a common basis matrix
2. different coefficient matrices

$$X_i = WH_i$$
$$W \geq 0, H_i \geq 0, i = 1,2,3$$



Outline

ARTICLES

OPEN

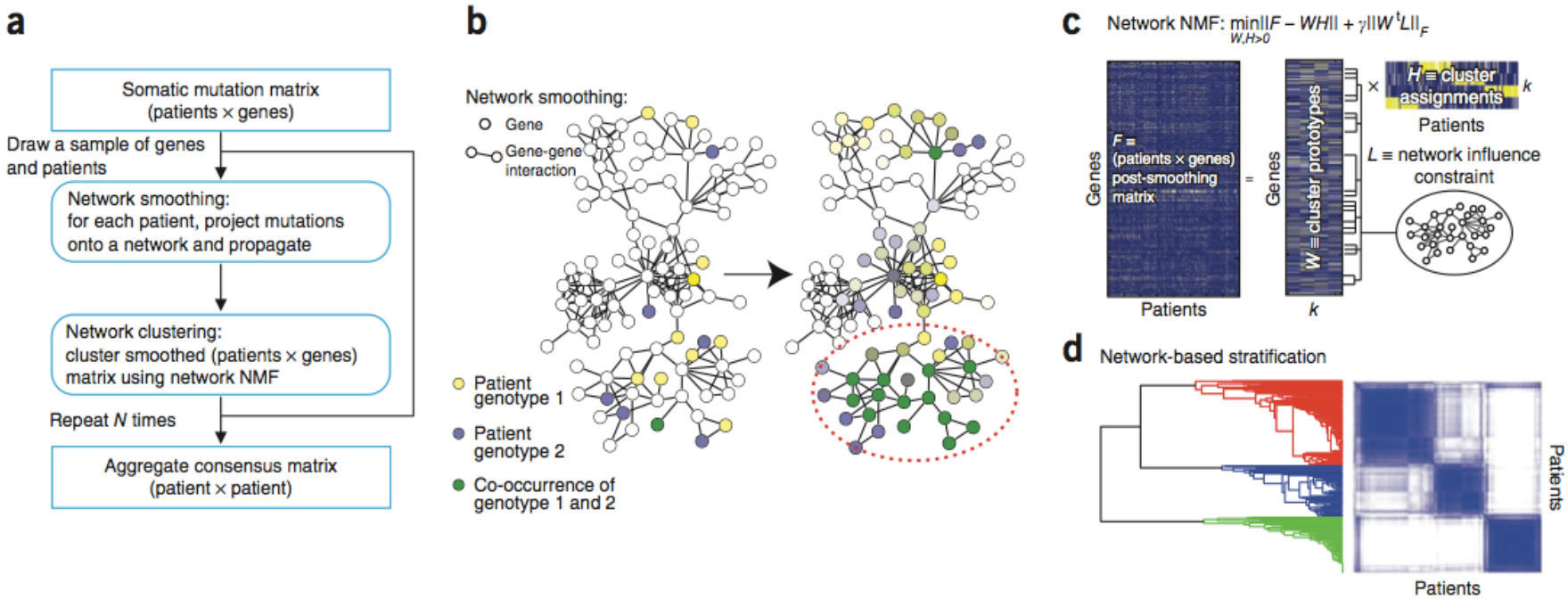
Network-based stratification of tumor mutations

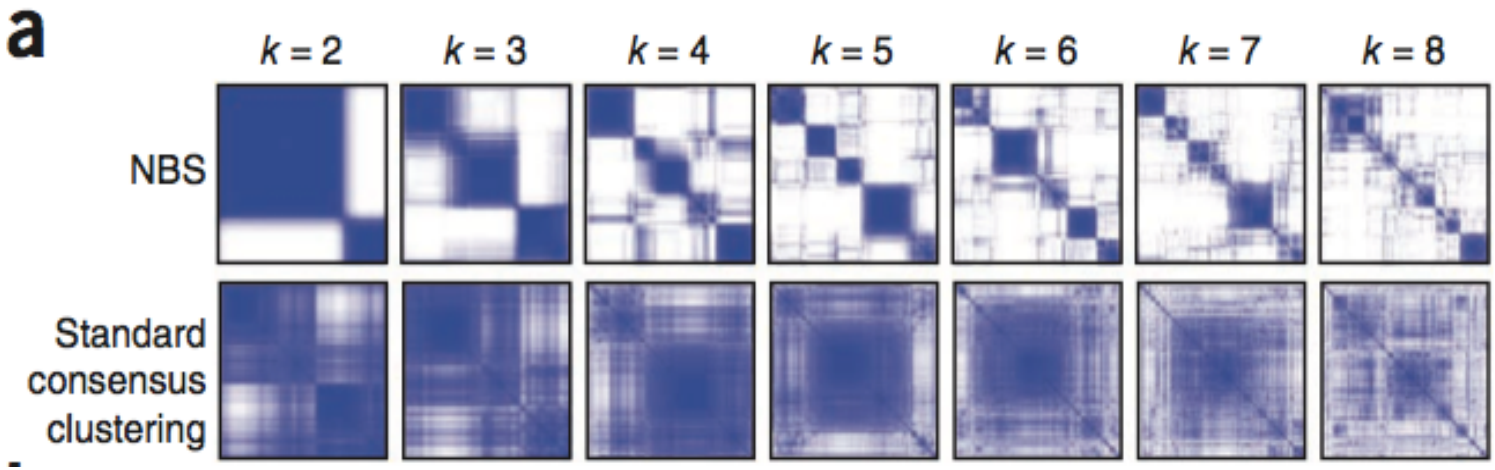
Matan Hofree[1], John P Shen[2], Hannah Carter[2], Andrew Gross[3] & Trey Ideker[1–3]

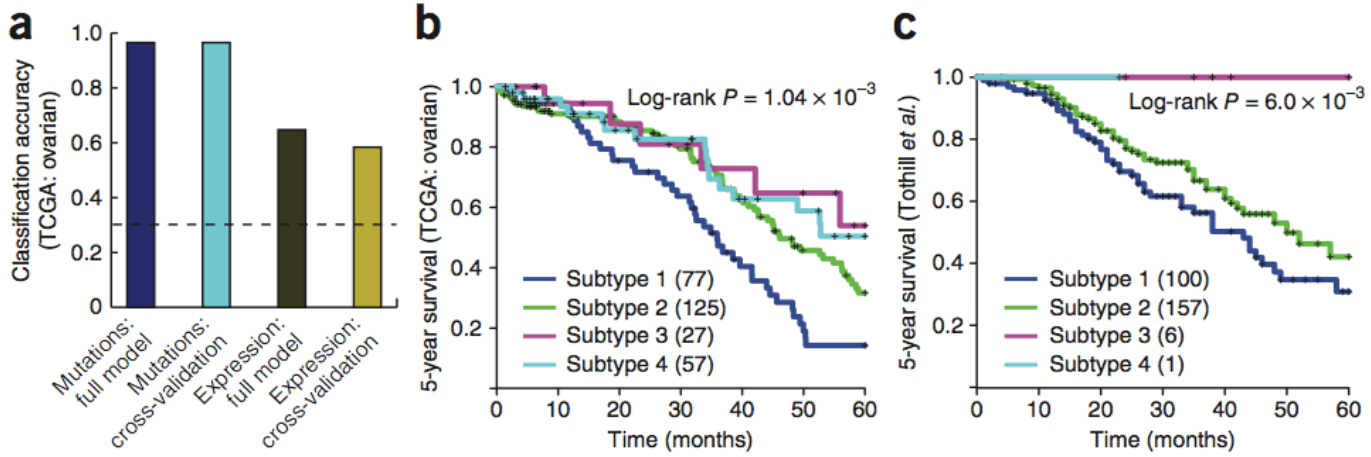
Overview

overview

Combine genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes




a
k = 2
k = 3
k = 4
k = 5
k = 6
k = 7
k = 8
NBS
Standard consensus clustering


a
Classification accuracy (TCGA: ovarian)
1.0
0.8
0.6
0.4
0.2
0
Mutations: full model
Mutations: cross-validation
Expression: full model
Expression: cross-validation
b
5-year survival (TCGA: ovarian)
Log-rank P = 1.04 × 10−3
Subtype 1 (77)
Subtype 2 (125)
Subtype 3 (27)
Subtype 4 (57)
0 10 20 30 40 50 60
Time (months)
c
5-year survival (Tothill et al.)
Log-rank P = 6.0 × 10−3
Subtype 1 (100)
Subtype 2 (157)
Subtype 3 (6)
Subtype 4 (1)
0 10 20 30 40 50 60
Time (months)

Discussion

1. capture organization and structure within the data
2. learn parts-based representations and cause more reasonable intepretation
3. integrate multiple data source to detect co-module
4. effective dimensionality reduction