A novel pathway-based distance score enhances assessment of disease heterogeneity in gene expression

Yunqing Liu
Yale University School of Public Health, yunqing.liu@yale.edu

Xiting Yan
Yale University School of Medicine, xiting.yan@yale.edu

Yale School of Medicine

A novel pathway-based distance score enhances assessment of disease heterogeneity in gene expression

Yunqing Liu[1], Jenny Lee[1], Anqi Liang[1], Hongyu Zhao[1], Geoffrey L. Chupp[2], Xiting Yan[1,2]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA [2]Center for Pulmonary Personalized Medicine (P2MED), Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, CT, USA.

PCCSM

P2MED

BACKGROUND

- Unsupervised clustering of patients using gene expression data is popularly used to study disease heterogeneity.
- Traditional Euclidean distance may not be efficient at discriminating the biological differences between samples due to the high noise to signal ratio in gene expression data.
- Distance scores defined based on pre-defined pathways instead of individual genes may help reduce the noise to signal ratio and integrate prior biological knowledge.
- We assume that differences in the expression levels of genes from the same pathway are more predictive of the biological differences compared to standard approach and if integrated into clustering analysis, will enhance the robustness and accuracy of the clustering results.

METHODS

- Pre-defined biological pathways $\{P_k: k = 1,2,\cdots,K\}$ were downloaded from KEGG, where P_k is the set of genes in pathway k.
- The N patients were clustered using expression levels of genes from each pathway separately based on a Gaussian Mixture Model. Let $C_k = (c_1^k, c_2^k, \cdots, c_N^k)$ be the clustering results of the patients using pathway P_k, in which c_j^k is an integer indicating which cluster the patient j is assigned to by pathway P_k.
- The clustering results across all the pathways were summarized into a pathway based distance score defined as follows. The distance between patient j_1 and j_2 is calculated as

$$d(j_1, j_2) = \frac{\#\{k: c_{j_1}^k \neq c_{j_2}^k, m_k > 1\}}{\#\{k: m_k > 1\}}$$

where m_k is the total number of patient clusters identified using pathway P_k and $\#\{\cdot\}$ is the size of the set $\{\cdot\}$.

- To demonstrate our method, we simulate gene expression data from 120 patients that belong to 3 groups with 40 patients per group. The expression levels of genes from pathway P_k are simulated from the following distribution:

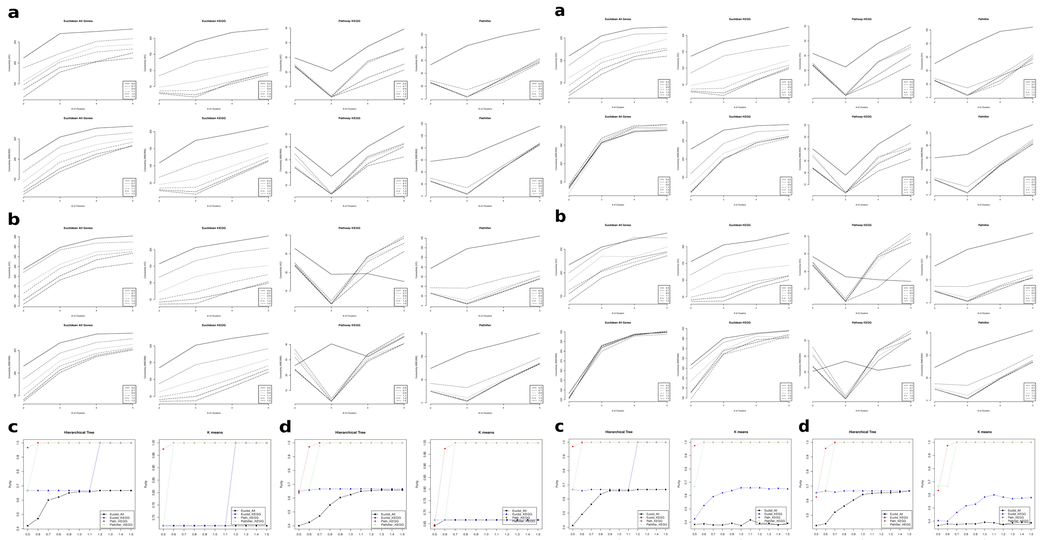
$$\begin{pmatrix} G_{i\Omega_k} \\ G_{i\overline{\Omega_k}} \end{pmatrix} \sim Gaussian\left(\begin{pmatrix} \mu_{C_i} \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_0 & \rho\Pi \\ \rho\Pi & \Sigma_1 \end{pmatrix}\right)$$

where Ω_k is the set of genes from pathway P_k that are differentially expressed across the 3 groups, $G_{i\Omega_k}$ is the vector of expression levels of the genes from Ω_k in subject i, C_i indicates which group subject i belongs to, $\mu_{C_i} = \begin{cases} -\delta, if\ C_i = 1 \\ 0, if\ C_i = 2 \\ \delta, if\ C_i = 3 \end{cases}$, $\Sigma_0 = \begin{bmatrix} \sigma^2 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & \sigma^2 \end{bmatrix}$, $\sigma^2 = 1 + \frac{2\delta^2}{3}$, $\Sigma_1 = \begin{bmatrix} B\sigma^2 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & B\sigma^2 \end{bmatrix}$ and $\Pi = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$.

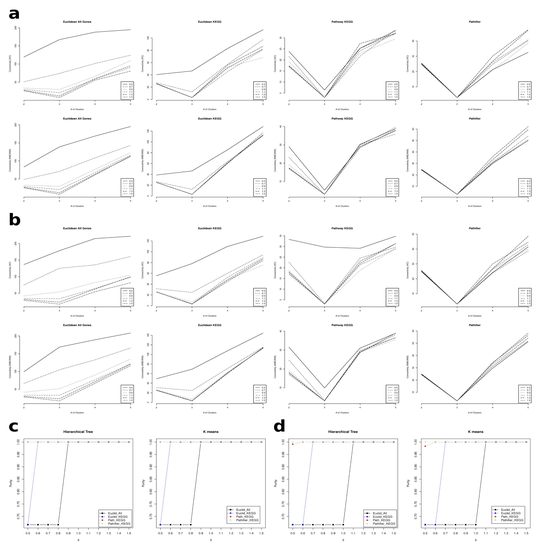
- Meaning of of simulation parameters:
  - B represents the background noise of genes that are not differentially expressed.
  - δ represents the amount of differences in the gene expression profiles between the 3 groups.
  - ρ represents the correlation coefficient between genes in the same pathway.
  - p_G represents the proportion of genes in a pathway that are differentially expressed.
- Our method was compared to Pathifier and the traditional Euclidean distance.

RESULTS

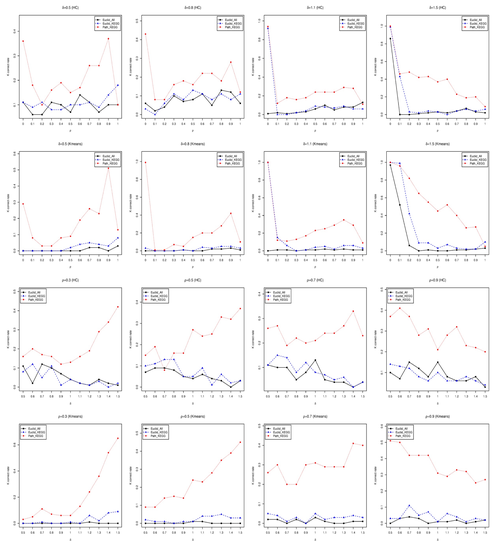
Simulation (low dimension): Connectivity criteria comparison ($\rho = 0$) for B=1 and 3.



Simulation (high dimension): Connectivity criteria ($\rho = 0$) for B=1 and 3.



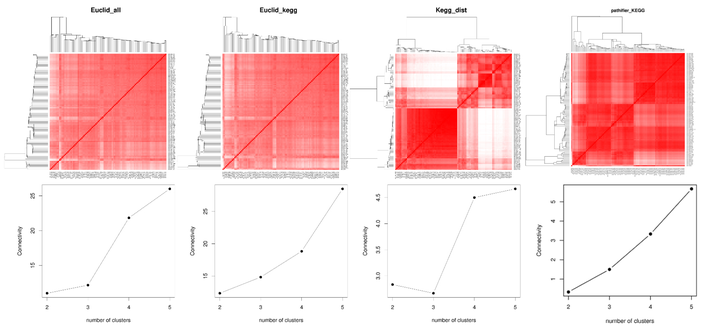
Accuracy of identifying the true number of clusters for $\rho \neq 0$.



Accuracy rate of identifying the true number of clusters for $\rho = 0, B = 1$ and $p_G = 0.2$.

δ		0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
HC	Euclid All	13%	13%	10%	8%	7%	7%	2%	6%	1%	10%	7%
	Euclid KEGG	6%	8%	2%	3%	4%	2%	3%	19%	35%	55%	72%
	PBS KEGG	22%	37%	34%	38%	45%	61%	75%	89%	98%	99%	100%
Kmeans	Euclid All	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Euclid KEGG	0%	0%	0%	0%	0%	0%	3%	19%	39%	54%	77%
	PBS KEGG	19%	50%	77%	92%	97%	97%	100%	100%	100%	99%	100%

Distance matrices comparison on asthma gene expression data



Phenotypic and physiologic characteristics of the identified clusters

	Euclid_all	Euclid_KEGG	KEGG_dist	Pathifier_KEGG
Age at Visit (years)	0.65	0.37	0.32	0.28
Gender	0.02*	0.14	0.58	0.28
History of Atopy - N (%)	**0.89**	**0.2**	**0.02**	**0.62**
Age of Symptom Onset	0.55	0.25	0.17	0.62
Disease Duration (years)	0.98	0.9	0.67	0.38
History of Hospitalization - N (%)	**0.21**	**0.77**	**0.04**	**1.00**
History of Intubations - N (%)	**0.14**	**0.12**	**0.05**	**0.04**
OCS tapers in past year- N (%)	0.65	1.00	0.67	0.83
ACT Score	0.25	0.41	0.22	0.56
FEV1- % of predicted value				
Pre β_2 agonist use	**0.04**	**0.02**	**0.02**	**0.04**
Post β_2 agonist use	0.06	0.05*	0.06	0.06
FVC- % of predicted value				
Pre β_2 agonist use	**0.04**	**0.02**	**0.04**	**0.03**
Post β_2 agonist use	0.12	0.06	0.16	0.13
FEV1/FVC- % of predicted value				
Pre β_2 agonist use	0.23	0.46	0.13	0.41
Post β_2 agonist use	0.14	0.2	0.06	0.09
BDR (%)	**0.27**	**0.05**	**0.05**	**0.09**
FENO (ppb)	0.05*	0.54	0.27	0.40

CONCLUSIONS

- We have developed a novel distance to represent the biological difference between samples using gene expression data.
- The comparison of this distance score to the Euclidean distance showed a better performance in both identifying the true number of clusters and assigning the samples to the correct classes.
- The comparison of this score to Pathifier showed a better performance and robustness for pathways with a small number of genes.
- Ongoing work on using a regularized Gaussian Mixture Model for clustering using each pathway.

FUNDINGS

This work is supported by NIH grants GM059507 (Dr. Zhao); R01HL118346, UH2HL123876 (Dr. Chupp); K01HL125474, FAMRI Young Clinical Scientist Award 113,393 (Dr. Gomez); and the National Center for Advancing Translational Science (NCATS) grant UL1 TR000142, R21LM012884 (Dr. Yan).