Mining brain-wide gene expression data to identify imaging genomic modules via biclustering

Poster No:

1504

Submission Type:

Abstract Submission

Authors:

Jingxuan Bao[1], Mansu Kim[1], Xiaohui Yao[1], Trang Le[1], Patryk Orzechowski[1], Jingwen Yan[2], Andrew Saykin[3], Jason Moore[1], Li Shen[1]

Institutions:

[1]University of Pennsylvania, Philadelphia, PA, [2]Indiana University-Purdue University Indianapolis, Indianapolis, IN, [3]Indiana University, Indianapolis, IN

First Author:

Jingxuan Bao
University of Pennsylvania
Philadelphia, PA

Co-Author(s):

Mansu Kim
University of Pennsylvania
Philadelphia, PA
Xiaohui Yao
University of Pennsylvania
Philadelphia, PA
Trang Le
University of Pennsylvania
Philadelphia, PA
Patryk Orzechowski
University of Pennsylvania
Philadelphia, PA
Jingwen Yan
Indiana University-Purdue University Indianapolis
Indianapolis, IN
Andrew Saykin
Indiana University
Indianapolis, IN
Jason Moore
University of Pennsylvania
Philadelphia, PA
Li Shen
University of Pennsylvania
Philadelphia, PA

Introduction:

Allen Human Brain Atlas (AHBA) [1], a brain-wide genome-wide (BWGW) gene expression data set, is a natural connection between genome and brain. We previously proposed to identify meaningful sub-portions of AHBA,

called imaging genomic modules (IGMs), to capture local co-expression patterns across imaging and genomic domains, and then used IGMs to help mine high level imaging genetic associations [2]. Our prior method applied hierarchical clustering twice to partition genes and brain regions of interest (ROIs) separately, and had two limitations: (1) it could only identify grid-like non-overlapping biclusters; and (2) global correlations were used for clustering, which was not suitable for finding local co-expression patterns. In this work, we propose a new method to overcome these limitations.

Methods:

AHBA includes BWGW gene expression data from 3,702 distinct tissues (from 6 donors) that covered the entire brain. We downloaded the AHBA data preprocessed by [3], which mapped expression data to ROIs in the HCP-MMP atlas [4]. Our analyses focused on 180 ROIs in the left hemisphere. Starting from the gene-ROI data matrix, we applied three biclustering methods (Plaid [5], Xmotif [6], FABIA [7]) to the matrix, and compared those with two baseline clustering methods (k-means and hierarchical) applied to both genomic and brain dimensions. For IGMs with a large number of genes, we further performed k-means clustering again to the genomic dimension to obtain new IGMs with smaller sizes.

We evaluated the quality of the identified IGMs via comparing them with randomly selected modules of the same size. Since we hypothesize that co-expressed genes/ROIs tend to be functionally related, we encourage the selection of IGMs with high intra-module correlations. With this, we proposed 3 different IGM evaluation metrics: 1) intra-module correlation distribution, 2) a principal component analysis (PCA) score quantifying that fewer PC capturing more variance within a module implies a more correlated module, and 3) a non-negative matrix factorization (NMF) score quantifying that more correlated modules tend to have a lower rank NMF result. Also, we performed pathway analysis on the identified gene set using enrichR [8] and functional annotation on the identified ROI set.

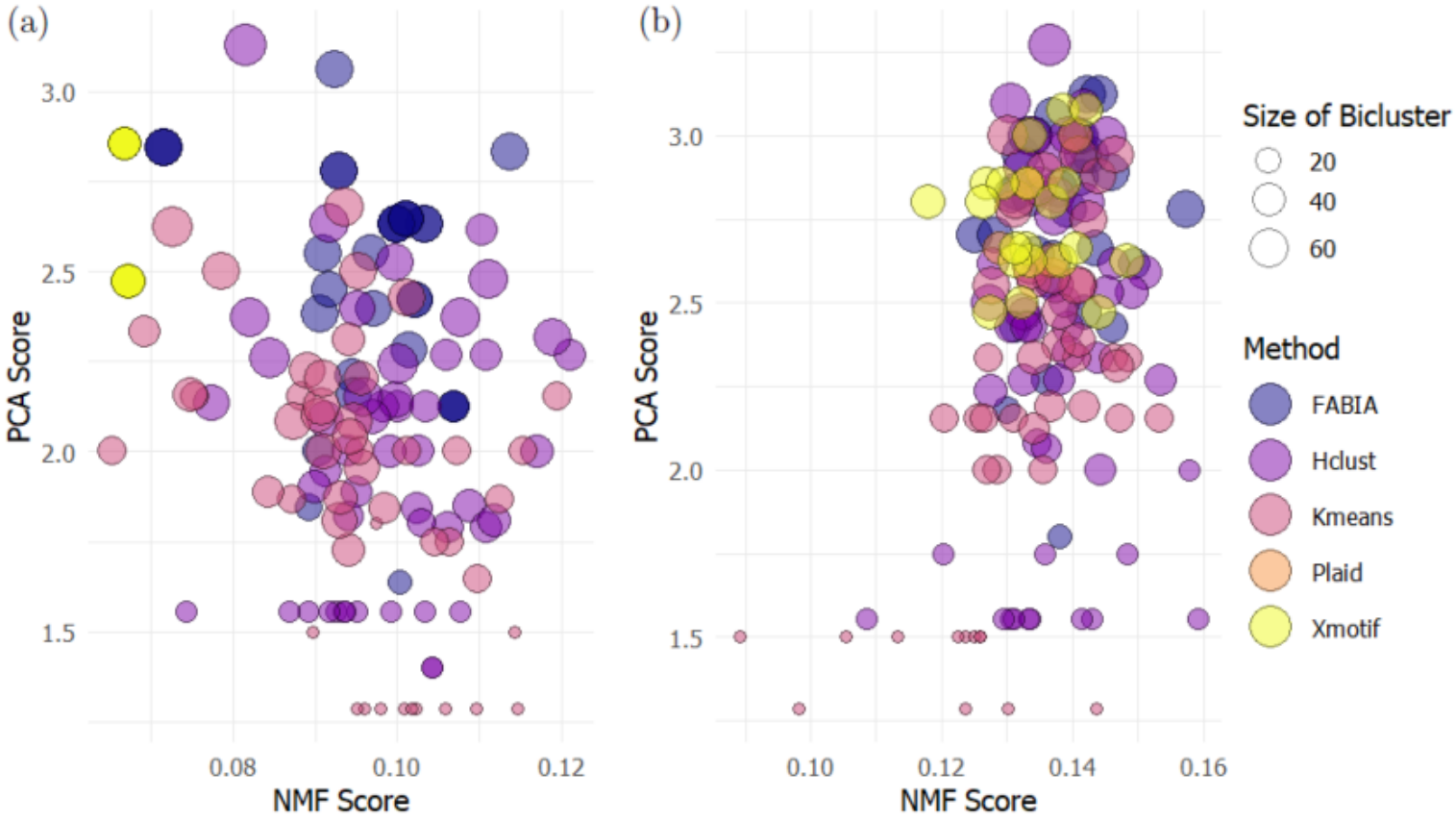
Results:

Figure 1.1 shows the correlation distribution results, where the identified IGMs show overall higher within-module correlations than the random modules with respect to both gene and ROI dimensions. Figure 1.2 shows the performance comparison using the PCA score (the larger the better) and NMF score (the smaller the better), where the identified IGMs also show better quality than random modules. From Figures 1.1 and 1.2, we also note that the three biclustering methods outperformed the two baseline clustering methods, and FABIA yielded the most promising results. Figure 2 shows pathway analysis results using KEGG_2019_Human database [9,10] and parcellation information from the HCP-MMP atlas [4] to provide potential biological interpretation for a few top identified IGMs.

Figure 1.1: **Summary Results of Identified Biclustering Modules.** The table summarized the main statistics for biclustering results obtained from different biclustering methods. The statistics include the number of identified biclustering modules ($\#Bic$), the $mean \pm sd$ for the number of identified genes ($\#Genes$) and ROIs ($\#ROIs$), and the $mean \pm sd$ for the absolute value of the mean of the correlation coefficients for genes ($Corr(G)$) and ROIs ($Corr(R)$).

Biclustering Results							
Method	$\#Bic$	$\#Genes$	$\#ROIs$	Identified Modules		Random Modules	
				$Corr(G)$	$Corr(R)$	$Corr(G)$	$Corr(R)$
K-means	48	46 ± 12	44 ± 27	0.39 ± 0.06	0.19 ± 0.03	0.27 ± 0.06	0.06 ± 0.24
Hierarchical	49	50 ± 11	44 ± 23	0.37 ± 0.10	0.19 ± 0.02	0.26 ± 0.05	0.05 ± 0.24
Plaid	4	46 ± 6	42 ± 0	0.55 ± 0.03	0.22 ± 0.03	0.25 ± 0.02	0.02 ± 0.24
Xmotif	20	46 ± 6	42 ± 0	0.55 ± 0.02	0.22 ± 0.03	0.24 ± 0.02	0.02 ± 0.23
FABIA	31	46 ± 10	53 ± 3	0.57 ± 0.12	0.20 ± 0.02	0.24 ± 0.01	0.01 ± 0.23

Figure 1.2: **Overall Comparison of PCA Score and NMF Score.** Figure (a) shows the comparison of PCA score and NMF score among the identified biclustering modules; Figure (b) shows the comparison of PCA score and NMF score among the randomly picked modules. Note that the upper left corner indicates better performance.



(https://files.aievolution.com/prd/hbm2101/abstracts/abs_1433/Figure1-new.png)

Figure 2.1: **Biological Evaluation (Pathway Analysis and Functional Annotation).** (a) shows the gene pathway analysis using KEGG_2019_Human database; (b) shows the human brain functional annotation using HCPMMP atlas. Every column of the heat map indicate one identified biclustering modules by the FABIA method where the color represents the $-log_{10}(p\text{-}value)$. All significant values are annotated.

(https://files.aievolution.com/prd/hbm2101/abstracts/abs_1433/Figure2-new.png)

Conclusions:

We proposed a new biclustering analysis to identify meaningful IGMs from AHBA. Our new method overcame previous limitations and was able to discover IGMs 1) with strong intra-module local co-expression patterns and 2) without grid-like nonoverlapping restriction. As a natural bridge between genome and brain, these identified IGMs can provide valuable information to guide the search for meaningful brain imaging genetic associations. With additional source of evidence at the transcriptomic level, the identified brain imaging genetic associations may be able to capture molecular effects from genetic determinants to gene expressions and to brain phenotypes, and thus are biologically meaningful and less likely to be false positives.

Genetics:

Genetic Modeling and Analysis Methods [2]
Transcriptomics [1]

Modeling and Analysis Methods:

Other Methods

Neuroinformatics and Data Sharing:

Brain Atlases

Keywords:

Computing
Data analysis
Design and Analysis
Informatics
Machine Learning
Modeling
Phenotype-Genotype
Statistical Methods

[1|2]Indicates the priority used for review

My abstract is being submitted as a Software Demonstration.

No

Please indicate below if your study was a "resting state" or "task-activation" study.

Other

Healthy subjects only or patients (note that patient studies may also involve healthy subjects):

Healthy subjects

Are you Internal Review Board (IRB) certified? Please note: Failure to have IRB, if applicable will lead to automatic rejection of abstract.

Yes

Was any human subjects research approved by the relevant Institutional Review Board or ethics panel? NOTE: Any human subjects studies without IRB approval will be automatically rejected.

Yes

Was any animal research approved by the relevant IACUC or other animal research panel? NOTE: Any animal studies without IACUC approval will be automatically rejected.

Not applicable

Please indicate which methods were used in your research:

Computational modeling
Other, Please specify - Whole genome whole brain gene expression map

Provide references using author date format

1. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391-399. doi:10.1038/nature11405
2. Yao X, Yan J, Kim S, et al. Two-dimensional enrichment analysis for mining high-level imaging genetic

associations. Brain Inform. 2017;4(1):27-37. doi:10.1007/s40708-016-0052-4
3. Arnatkeviciute A, Fulcher BD, Fornito A. A practical guide to linking brain-wide gene expression and neuroimaging data. Neuroimage. 2019;189:353-367. doi:10.1016/j.neuroimage.2019.01.011
4. Glasser MF, Sotiropoulos SN, Wilson JA, et al. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013;80:105-124. doi:10.1016/j.neuroimage.2013.04.127
5. Lazzeroni L, Owen A. PLAID MODELS FOR GENE EXPRESSION DATA. Statistica Sinica. 2002;12(1):61-86. http://www.jstor.org/stable/24307036.
6. Murali TM, Kasif S. Extracting conserved gene expression motifs from gene expression data. Pac Symp Biocomput. 2003:77-88. https://www.ncbi.nlm.nih.gov/pubmed/12603019. Published 2003/02/27.
7. Hochreiter S, Bodenhofer U, Heusel M, et al. FABIA: factor analysis for bicluster acquisition. Bioinformatics. 2010;26(12):1520-1527. doi:10.1093/bioinformatics/btq227
8. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90-97. doi:10.1093/nar/gkw377
9. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30. doi:10.1093/nar/28.1.27
10. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28(11):1947-1951. doi:10.1002/pro.3715