ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 20(7): 700-704 (2024)

www.bioinformation.net

Volume 20(7)

Research Article

Received June 1, 2024; Revised July 4, 2024; July 4, 2024, Published July 4, 2024

DOI: 10.6026/973206300200700

BIOINFORMATION 2022 Impact Factor (2023 release) is 1.9.

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Disclaimer:
The views and opinions expressed are those of the author(s) and do not reflect the views or opinions of Bioinformation and (or) its publisher Biomedical Informatics. Biomedical Informatics remains neutral and allows authors to specify their address and affiliation details including territory where required. Bioinformation provides a platform for scholarly communication of data and information to create knowledge in the Biological/Biomedical domain.

Edited by P Kangueane
Citation: Ryan *et al.* Bioinformation 20(7): 700-704 (2024)

Interpreting and visualizing pathway analyses using embedding representations with PAVER

William G Ryan V[1], Ali Sajid Imami[1], Hunter Eby[1], John Vergis[1], Xiaolu Zhang[2], Jarek Meller[3,4,5,6], Rammohan Shukla[7] & Robert McCullumsmith[1,8,9,*]

[1]Department of Neurosciences, College of Medicine and Life Sciences, University of Toledo, Toledo, OH, USA; [2]Department of Microbiology and Immunology, Louisiana State University Health Sciences Center, Shreveport, LA, USA; [3]Department of Environmental and Public Health Sciences, University of Cincinnati, Cincinnati, OH, USA; [4]Department of Computer Science, University of Cincinnati, Cincinnati, OH, USA; [5]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [6]Department of Informatics, Nicolaus Copernicus University, Toruń, Poland; [7]Department of Zoology & Physiology, College of Agriculture, Life Sciences and Natural Resources, University of Wyoming, Laramie, WY, USA; [8]Neurosciences Institute, ProMedica, Toledo, OH, USA; [9]Department of Psychiatry, College of Medicine and Life Sciences, University of Toledo, Toledo, OH, USA; *Corresponding author

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 20(7): 700-704 (2024)

Author contacts:
William G Ryan V - E-mail: William.Ryan2@rockets.utoledo.edu
Ali Sajid Imami - E-mail: Ali.Imami@rockets.utoledo.edu
Hunter Eby - E-mail: Hunter.Eby@rockets.utoledo.edu
John Vergis - E-mail: John.Vergis@utoledo.edu
Xiaolu Zhang - E-mail: xiaolu.zhang@lsuhs.edu
Jarek Meller - E-mail: mellerj@ucmail.uc.edu
Rammohan Shukla - E-mail: rshukla@uwyo.edu
Robert McCullumsmith - E-mail: robert.mccullumsmith@utoledo.edu; Phone: +1 205-789-0841 Fax: +1 419-383-3008

Abstract:
Omics studies use large-scale high-throughput data to explain changes underlying different traits or conditions. However, omics analysis often results in long lists of pathways that are difficult to interpret. Therefore, it is of interest to describe a tool named PAVER (Pathway Analysis Visualization with Embedding Representations) for large scale genomic analysis. PAVER curates similar pathways into groups, identifies the pathway most representative of each group, and provides publication-ready intuitive visualizations. PAVER clusters pathways defined by their vector embedding representations and then identifies the term most cosine similar to its respective cluster's average embedding. PAVER can integrate multiple pathway analyses, highlight relevant biological insights, and work with any pathway database.

Keywords: Biocuration, computational biology, gene expression profiling, R, systems biology, pathway analysis

Availability: PAVER with source code is made available at https://github.com/willgryan/PAVER. & The PAVER web application is available at https://cdrl.shinyapps.io/PAVER.

Background:
Multiomics, like transcriptomics, proteomics and kinomics, are used today in experimental biological research to study systems of disease and for precision medicine in clinical settings **[1, 2]**. The development of these technologies has outpaced researcher's expertise in analyzing data they collect **[3]**. This "data deluge" exceeds the capacity of human cognition **[4, 5]**. Analysis of omics is now a leading expense and bottleneck in most projects, limiting its translation from bench-to-bedside **[6-8]**. Pathway analysis has since become common to interpret high-throughput experiments and explain mechanisms of biological phenomena **[9]**. However, pathway analysis generally outputs lists of results too long to manually inspect **[10, 11]**. Various applications have been developed accordingly to summarize information from pathway analyses by selecting most representative terms (MRTs) - the key biological theme defining functionally related groups of pathways - using semantic similarity of Gene Ontology (GO) terms **[12-16]**. Semantic similarity measures closeness in meaning between GO terms for interpretation, gene clustering and disease-gene prediction **[17]**. These applications tend to lack complete interoperability beyond controlled ontologies like GO; restricting them from other pathway knowledge bases **[18-20]**. The growing volume of omics data indicates a need for novel ways of data management, like automated interpretation of omics results **[21, 22]**.

Modern AI is now being applied to biomarker discovery, survival prognosis and disease subtyping **[23, 24]**. Recently, embedding models that generate representations of biomedical ontologies have been developed for machine learning techniques like clustering and visualization **[25-27]**. Embeddings are numeric vector definitions of pathways that capture their meaning for use as a measure of semantic similarity **[26, 28]**. This allows for mathematics between words e.g., "Genome - Genes + Proteins = Proteome," where the meaning of different words can be averaged to capture their overall sentiment **[29]**. On biomedical corpora, embeddings can represent millions of words in hundreds of numerical dimensions **[30, 31]**. Representing the combined meaning of words with their average embedding in this way has been applied to biological prediction tasks **[32]**. Embedding models have also been used to define biological entities, like pathways, as the average embedding of their constituent gene members to predict protein-protein interactions **[33, 34]**. Here, we present PAVER, a novel method that extends this concept by using embedding representations to measure semantic similarity of pathways and identify MRTs in groups of related pathways (**Figure 1A**). The PAVER algorithm (**Figure 1B**) first hierarchically clusters pathway embedding's. Pathway embedding's then averaged for each cluster to capture its overall meaning into a single numerical representation. The MRT is finally selected by determining which pathway is most cosine similar to its respective cluster's average embedding. This allows PAVER to curate long lists of pathways into related groups and identify the pathway most representative of each group. PAVER is implemented in a freely available R programming language software package and web application for researchers to integrate, interpret and visualize common pathway analysis outputs.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 20(7): 700-704 (2024)

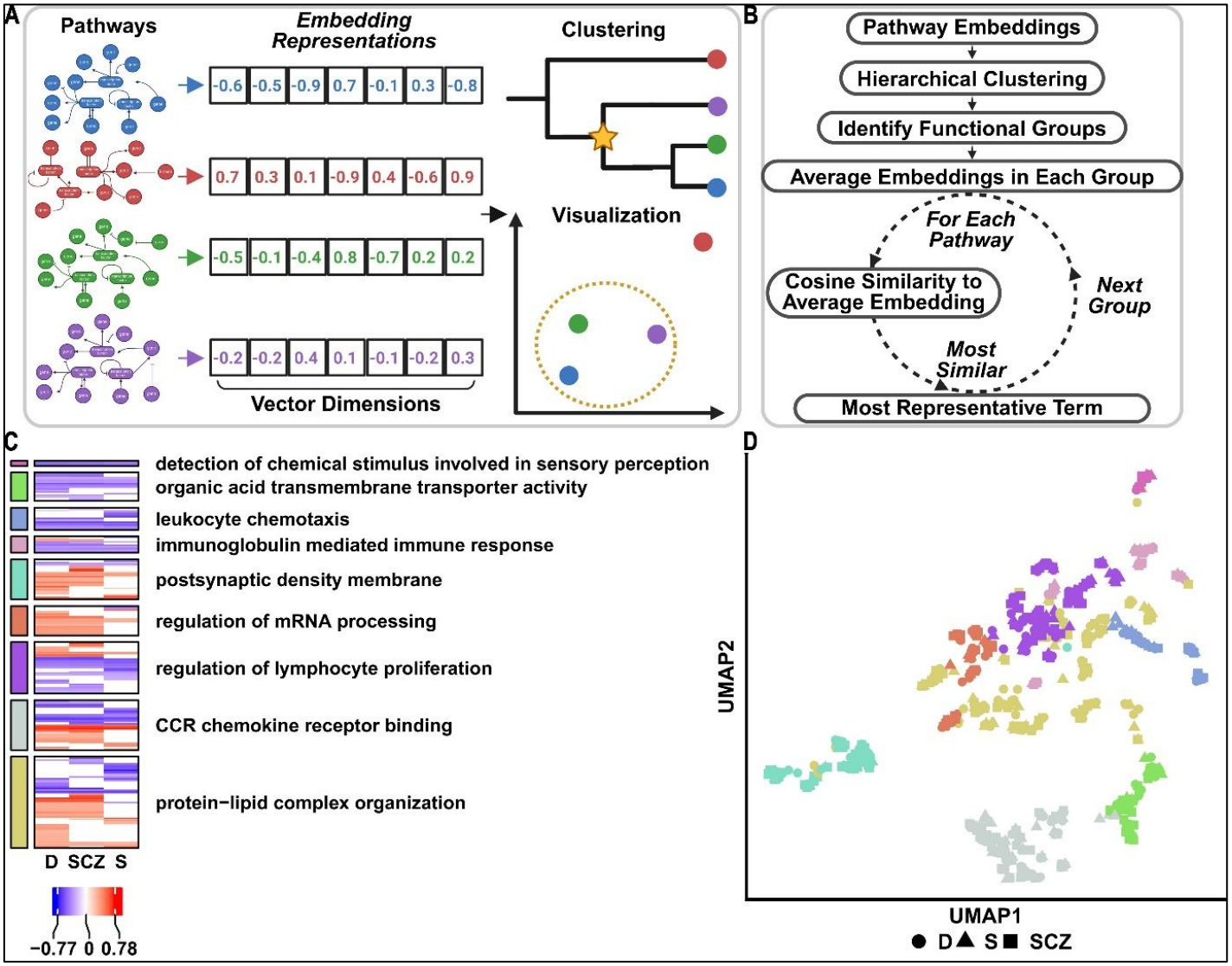

Figure 1: PAVER uses numerical representations of biological pathways to identify functionally related clusters
(A) Conceptual overview of the PAVER method implemented in an R programming language software package and web application. Precomputed embedding representations of biological pathways are used for clustering and visualization to aid interpretation of pathway analyses. (B) Diagram of the underlying PAVER algorithm. PAVER is a novel method to select MRTs from groups of functionally related pathways by averaging their embeddings and determining which individual pathway is most cosine similar to its respective group's average. (C) A heatmap generated by the PAVER R package showing uniquely colored-coded clusters of pathways and their identified MRTs from a previously manually interpreted pathway analysis that delineated deep (D), superficial (S), or combined (SCZ) cortical lamina neurons in a bulk RNAseq study of postmortem chronic schizophrenia brain. Legend shows enrichment score from GSEA [40] (D) A scatterplot generated by the PAVER R package showing the 2D computed UMAP of the pathway embeddings. Points show GO terms. Shape indicates respective pathway analysis. Color shows cluster membership for each pathway. MRT: Most Representative Term, GSEA: Gene-set enrichment analysis, 2D: two-dimension, UMAP: Uniform Manifold Approximation and Projection

Input and Output:
PAVER requires two inputs: pathway analysis results and pre-computed pathway embedding's. Pathway analysis results are expected to be a wide-format table where the first column contains pathway identifiers (e.g. GO: 0005739, hsa04512, WP4562, etc.) and the following columns contain their respective enrichment metrics (e.g. p-value, enrichment score, combined score, etc.) returned from tools like Enrichr or gene-set enrichment analysis **[35, 36]**. PAVER works generally with any set of pre-computed embedding's. PAVER provides precomputed pathway embedding's using the recent anc2vec embedding model of GO. **[25]** PAVER also provides

precomputed pathway embedding's for GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) using the recent "text-embedding-3-large" embedding model provided by Open AI. These Open AI models have been shown to link biomedical concepts like relationships between diseases, genes and epidemiology **[37-39]**. We created multi-lined word strings for submission to the embedding model by concatenating each GO term's ID, sub-ontology name, and definition or each KEGG pathway's entry, name, description and class. To demonstrate the utility of PAVER, we applied it to previously manually interpreted pathway analysis results to identify MRTs that delineate deep versus superficial cortical lamina neuron function in a bulk RNAseq study of postmortem chronic schizophrenia brain. **[40]** PAVER identified MRTs like *detection of chemical stimulus involved in sensory perception*, *postsynaptic density membrane*, and *CCR chemokine receptor binding* that closely mirrored manual curation, like *sensory system*, *synapse*, and *cytokine immunity*, and provided intuitive heat map-based (**Figure 1C**) and scatterplot-based (**Figure 1D**) visualizations. Notably, PAVER performed this curation and visualization task more quickly than could be achieved manually without.

Caveats and Future Development:

PAVER provides a novel method for summarization of biological pathways defined by their embedding representations. However, PAVER assumes the input pathway analysis was properly performed **[10]**. PAVER also requires that embedding representations are pre-computed. PAVER's proof-of-concept has previously been used in a number of studies to aid in the interpretation of pathway analyses and helps explain mechanisms underlying different disorders and diseases **[41-46]**. We plan to further increase the utility of PAVER with additional visualizations and pre-computed pathway embedding's for other pathway databases. We hope PAVER will continue to be a valuable resource to help researchers extract actionable insights from their pathway analyses. The PAVER R package is licensed under the GNU General Public License v3.0.

Declarations:

The authors declare that shinyapps.io is a hosting service provided by the public benefit corporation Posit. Posit has an excellent reputation for ensuring the up time of their hosted applications. Hence, the URL and application are both sustainable in the long term. We have previous experience using this service to host another application which has been available without interruption for more than five years. Further, University of Toledo IT security policy prevents us from using the utoledo.edu domain to host applications.

Acknowledgements:

This work was supported by NIH NIGMS T32-G-RISE grant number 1T32GM144873-01, NIH NIMH grant number R01MH107487, NIH NIMH grant number R01MH121102, and NIH NIA grant number R01AG057598.

References:

[1] Veenstra TD. *Proteomics* 2021 **21**:3-4 [PMID:33320441].
[2] Herr TM *et al*. *Journal of Pathology Informatics* 2015 **6**:46 [PMID:26430534].
[3] Denecker T & Lelandais G. *Methods Mol Biol* 2022 **2477**:457471 [PMID:35524132].
[4] Bell G *et al*. *Science* 2009 **323**:1297 [PMID:19265007].
[5] Stead WW *et al*. *Academic Medicine* 2011 **86**:429 [PMID:20711055].
[6] Krassowski M *et al*. *Front Genet* 2020 **11**:610798 [PMID:33362867].
[7] D'Adamo GL *et al*. *Immunol Cell Biol* 2021 **99**:168 [PMID:32924178].
[8] Sboner A *et al*. *Genome Biol* 2011 **12**:125 [PMID:21867570].
[9] García-Campos MA *et al*. *Front Physiol* 2015 **6**:383 [PMID:26733877].
[10] Chicco D & Agapito G. *PLoS Comput Biol* 2022 **18**:e1010348 [PMID:35951505].
[11] Supek F & Škunca N. *The Gene Ontology Handbook* 2017 **1446**:207220 [PMID:27812945].
[12] Wang G *et al*. *BMC Bioinformatics* 2020 **21**:139 [PMID:32272889].
[13] Ewing E *et al*. *BMC Bioinformatics* 2020 **21**:443 [PMID:33028195].
[14] Supek F *et al*. *PLoS One* 2011 **6**:e21800 [PMID:21789182].
[15] Yu G *et al*. *Omics* 2012 **16**:284 [PMID:22455463].
[16] Reijnders MJ & Waterhouse RM. *Frontiers in Bioinformatics* 2021 **1**:638255 [PMID:36303779].
[17] Pesquita C *et al*. *PLOS Computational Biology* 2009 **5**:c1000443 [PMID:19649320].
[18] Zhao C & Wang Z. *Scientific Reports* 2018 **8**:15107 [PMID:30305653].
[19] Gan M *et al*. *The Scientific World Journal* 2013 **2013**:793091 [PMID:23533360].
[20] Galeota E *et al*. *Scientific Reports* 2020 **10**:703 [PMID:31959844].
[21] Perez-Riverol Y *et al*. *Nat Commun* 2019 **10**:3512 [PMID:31383865].
[22] Wilkinson MD *et al*. *Sci Data* 2016 **3**:160018 [PMID:26978244].
[23] https://pubmed.ncbi.nlm.nih.gov/31815397/
[24] MacEachern SJ & Forkert ND. *Genome* 2021 **64**:416 [PMID:33091314].
[25] Edera AA *et al*. *Briefings in Bioinformatics* 2022 **23**:bbac003 [PMID:35136916].
[26] Kulmanov M *et al*. *Briefings in Bioinformatics* 2021 **22**:bbaa19 [PMID:33049044].
[27] Duong D *et al*. *Journal of Computational Biology* 2019 **26**:38 [PMID:30383443].
[28] Lerman G & Shakhnovich BE. *Proceedings of the National Academy of Sciences* 2007 **104**:11334 [PMID:17595300].
[29] Mikolov T *et al*. *arXiv preprint arXiv:1301.3781* 2013 [doi:10.48550/arXiv.1301.3781].
[30] Major V *et al*. *AMIA Annu Symp Proc* 2018 **2018**:14051414 [PMID:30815185].

[31] Chiu B *et al*. *Proceedings of the 15th workshop on biomedical natural language processing* 2016 **W16**:2922 [doi:10.18653/v1/W16-2922].
[32] Ofer D *et al*. *Computational and Structural Biotechnology Journal* 2021 **19**:1750 [PMID:33897979].
[33] Xenos A *et al*. *Bioinformatics* 2021 **37**:3839 [PMID:34213534].
[34] Asgari E & Mofrad MRK. *PLOS ONE* 2015 **10**:c0141287 [PMID:26555596].
[35] Subramanian A *et al*. *Proc Natl Acad Sci U S A* 2005 **102**:15545 [PMID:16199517].
[36] Kuleshov MV *et al*. *Nucleic Acids Res* 2016 **44**:W90 [PMID:27141961].
[37] Brown T *et al*. *Advances in neural information processing systems* 2020 **33**:4 [doi:10.48550/arXiv.2005.14165].
[38] Wang Q *et al*. *arXiv preprint arXiv:2307.01137* 2023 [doi:10.48550/arXiv.2307.01137].
[39] Wang D-Q *et al*. *MedComm – Future Medicine* 2023 **2**:e43 [doi:10.1002/mef2.43].
[40] Wu X *et al*. *Mol Psychiatry* 2021 **26**:7699 [PMID:34272489].
[41] Curtis MA *et al*. *bioRxiv* 2024 [PMID:37961675].
[42] Nguyen JH *et al*. *bioRxiv* 2024 [PMID:37745438].
[43] Hu Y *et al*. *Translation: The University of Toledo Journal of Medical Sciences* 2024 **12**:e1 [doi:10.46570/utjms.vol12-2024-823].
[44] O'Donovan S *et al*. *Translation: The University of Toledo Journal of Medical Sciences* 2024 **12**:1 [doi:10.46570/utjms.vol11-2023-822].
[45] Ryan W. *Zenodo* 2023 [doi:10.5281/zenodo.8156248].
[46] Ryan W *et al*. *Translation: The University of Toledo Journal of Medical Sciences* 2023 **11**:e1 [doi:10.46570/utjms.vol11-2023-906].

Note: This supersedes the previous version where page numbers for several references are corrected
September 20, 2024