

Построение однозначной функциональной аннотации генов с помощью метаанализа на основе графов

Бузанов Г.^{1, 2*}, Макеев В.¹

¹ Институт общей генетики РАН, Москва, Россия

² Московский физико-технический институт, Москва, Россия

* gleb.buzanov@phystech.edu

Ключевые слова: кластеризация генов; графовые алгоритмы; функциональная аннотация; Enrichr; OpenOrd; функции генов

Мотивация и цель: Исследования GWAS обычно дают ограниченное количество маркерных генов, недостаточно большое для проведения анализа, подобного GSEA. В этой работе мы предлагаем графовый алгоритм, который присваивает каждому гену одну функцию, обеспечивая тем самым однозначную функциональную аннотацию. Для этого большое количество функциональных сигнатур объединяются в единый взвешенный граф, который разбивается на кластеры. Для каждого маркерного гена формируется ограниченный набор генов, которые коэкспрессируются с маркерным геном в большом количестве исследований по профайлингу экспрессии и могут рассматриваться как первичные кандидаты на участие в одних и тех же процессах формирования фенотипа.

Методы и алгоритмы: Основным алгоритмом в работе является алгоритм построения графовой развертки OpenOrd, а также алгоритм K-Means для автоматизации процесса построения кластеризации. Результирующие кластеризации сравнивались по метрикам ассортативности и по мере Жаккара.

Результаты: Полученный алгоритм продемонстрировал лучшее качество кластеризации среди применимых к этой задаче алгоритмов кластеризации, таких как HDBSCAN, k-medoids и др., на выборках размером до 5000 генов. Кроме того, в отличие от других алгоритмов кластеризации, OpenOrd подходит и для более широких выборок, в результате чего возможна кластеризация полного набора из ~20 000 генов, доступных в базах Enrichr. Мы использовали Metascape для аннотирования генов, для выявления доказательств, подтверждающих рациональность кластеризации, и для получения информации о релевантности кластера.

В дальнейшем полученная кластеризация была заново агрегирована алгоритмом для получения иерархической структуры кластеризации. В результате двух дополнительных итераций было получено три слоя кластеризации. На менее детализированных слоях неизбежно возникает проблема слияния слабосвязанных генов в гиперкластеры. Учитывая, что иерархическая многослойная кластеризация логически выполняется на обширных наборах данных с большим количеством генов, можно сделать вывод, что такая организация данных оказывается неэффективной в данной конкретной задаче.

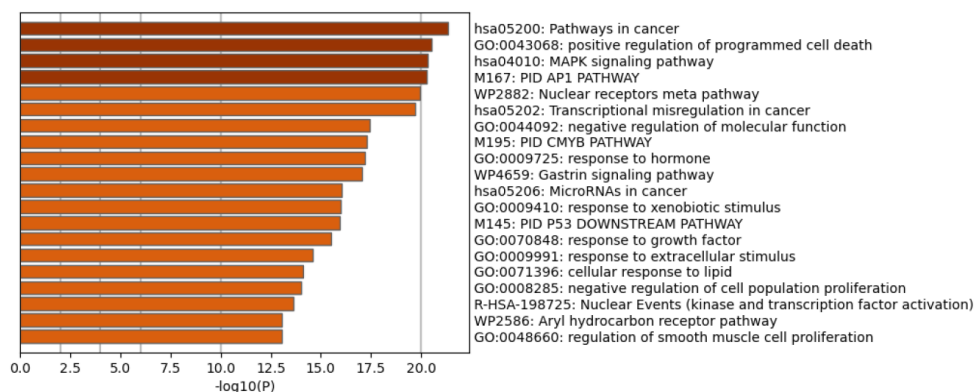


Figure 8: TP53 cluster analysis

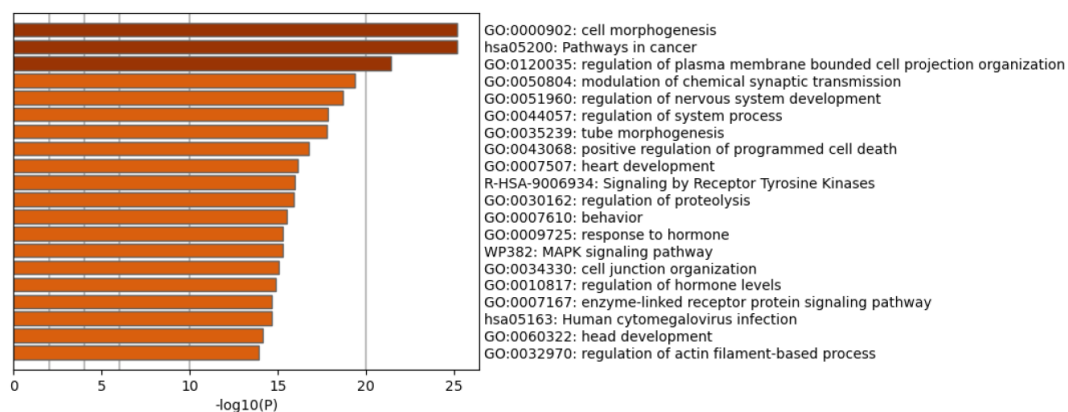


Figure 9: MYH14 cluster analysis

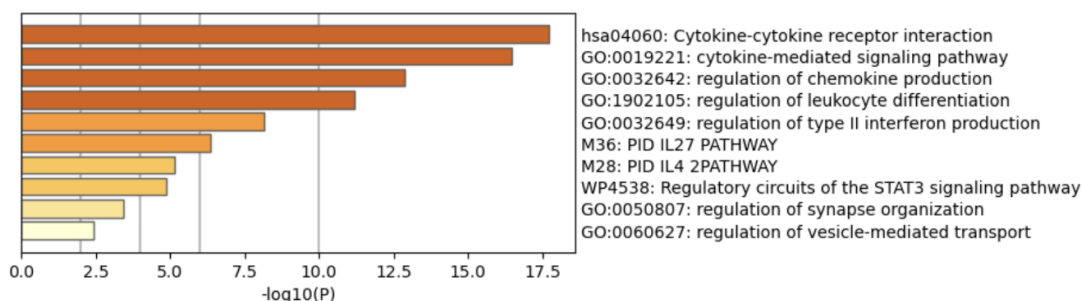


Figure 10: IL6 cluster analysis

Constructing an exclusive gene functional annotation with the help of a graph-based metaanalysis

Buzanov G.^{1, 2*}, Makeev V.¹

¹ Vavilov Institute of General Genetics, RAS, Moscow, Russia

² Moscow Institute of Physics and Technology, Moscow, Russia

* gleb.buzanov@phystech.edu

Key words: gene clustering; graph-based algorithm; functional annotation; Enrichr database; OpenOrd; gene functions

Motivation and Aim: GWAS studies usually produce a limited numbers of marker genes, not large enough to provide GSEA-like analysis. Here we suggest a graph based algorithms that assigns a single function to each gene, thus providing an exclusive functional annotation. To this end a large number of functional signatures are combined into a single weighted graph, which is segregated into cliques. For each marker gene a limited gene set is formed that are coexpressed with the marker gene in the large number of co-expression studies and can be considered as a primary candidates for taking part in the same processes of phenotype formation.

Methods and Algorithms: The main algorithm in the work is the OpenOrd graph scan algorithm, as well as the K-Means algorithm for automating the clustering construction process. The resulting clusterings were compared using assortativity metrics and the Jaccard measure.

Results: The resulting algorithm demonstrated the best quality of clustering among clustering algorithms applicable to this problem, such as HDBSCAN, k-medoids, etc. on samples of up to 5,000 genes. In addition, unlike other clustering algorithms, OpenOrd is also suitable for wider samples, as a result of which it is possible to cluster the entire set of about 20,000 genes available in the Enrichr databases. We used Metascape to annotate genes, to identify evidence supporting the rationality of clustering, and to obtain information about cluster relevance.

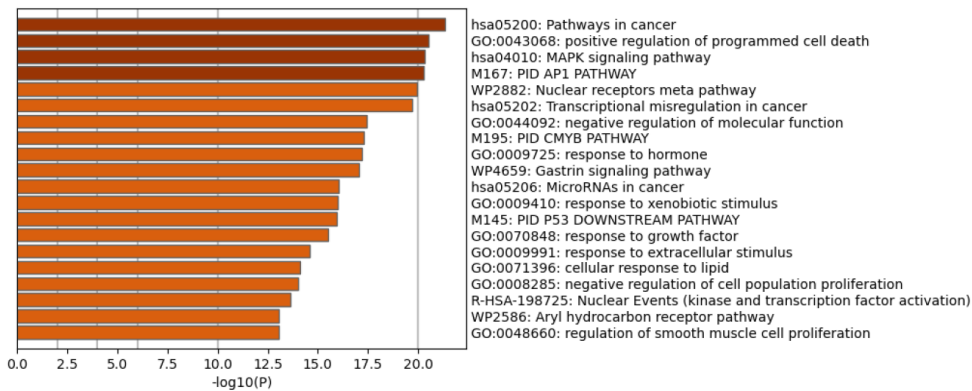


Figure 8: TP53 cluster analysis

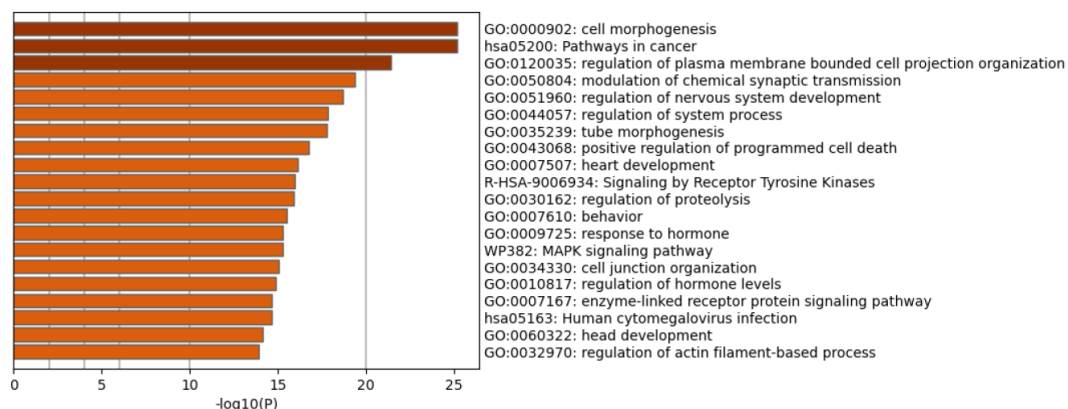


Figure 9: MYH14 cluster analysis

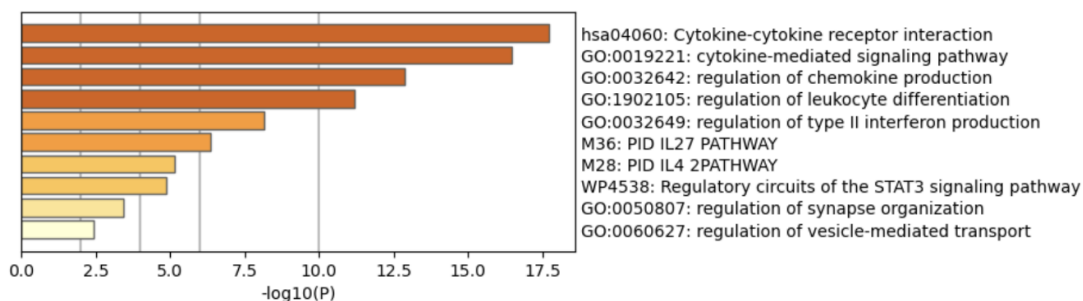


Figure 10: IL6 cluster analysis

The resulting clustering was then re-aggregated by the algorithm to obtain a hierarchical clustering structure. As a result of two additional iterations, three clustering layers were obtained. At less detailed layers, the problem of merging loosely connected genes into hyperclusters inevitably arises. Given that hierarchical multilayer clustering logically works on large datasets with large numbers of genes, it can be concluded that such data organization is ineffective for this particular task.

Список литературы/References

1. Ashburner M. et al. Gene ontology: tool for the unification of biology. *Nat Genet.* 2000;25(1):25-29
2. Ashworth A., Lord C.J., Reis-Filho J.S. Genetic interactions in cancer progression and treatment. *Cell.* 2011;145(1):30-38
3. Cano-Gamez E., Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet.* 2020;11:424
4. Chen E.Y., Tan C.M., Kou Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics.* 2013;14:128
5. Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* 2001;11(8):1425-1433
6. Harris M.A., Clark J., Ireland A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004;32(Database issue):D258-D261

7. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 2019;47(D1):D330-D338
8. Gu J., Dai J., Lu H., Zhao H. Comprehensive Analysis of Ubiquitously Expressed Genes in Humans from A Data-driven Perspective. *Genomics Proteomics Bioinformatics.* 2023;21(1):164-176
9. Hochane M., van den Berg P.R., Fan X. et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLoS Biol.* 2019;17(2):e3000152
10. Kuleshov M.V., Jones M.R., Rouillard A.D. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016;44(W1):W90-W97
11. Li X., Guo C., Ahmad S. et al. Systematic Analysis of MYB Family Genes in Potato and Their Multiple Roles in Development and Stress Responses. *Biomolecules.* 2019;9(8):317
12. Mar'n S., Brown W.M., Klavans R., Boyack K. OpenOrd: an open-source toolbox for large graph layout. In: *Visualization and Data Analysis.* 2011;7868:786806
13. Culhane A.C., Schwarzl T., Sultana R. et al. GeneSigDB – a curated database of gene expression signatures. *Nucleic Acids Res.* 2010;38(Database issue):D716-D725
14. Liberzon A., Subramanian A., Pinchback R. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27(12):1739-1740