НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2016 Том 16 № 5 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2016 Vol. 16 No 5 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.021

АЛГОРИТМ КУМУЛЯТИВНОГО ВЫЧИСЛЕНИЯ СТАТИСТИКИ ПРЕДСТАВЛЕННОСТИ НАБОРА ГЕНОВ

А.А. Сергушичев[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, alserg@rain.ifmo.ru
Адрес для переписки: alserg@rain.ifmo.ru

Информация о статье
Поступила в редакцию 08.07.16, принята к печати 10.08.16
doi: 10.17586/2226-1494-2016-16-5-956-959
Язык статьи – русский
Ссылка для цитирования: Сергушичев А.А. Алгоритм кумулятивного вычисления статистики представленности набора генов // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 5. С. 956–-959. doi: 10.17586/2226-1494-2016-16-5-956-959

Аннотация
Исследованы методы анализа представленности наборов генов, широко применяемые для анализа экспрессии генов. Рассмотрена задача кумулятивного вычисления статистики представленности. Для решения этой задачи предложен алгоритм, основанный на применении корневой эвристики. Найдена асимптотическая оценка на время работы алгоритма. Практическая реализация предложенного алгоритма показала ускорение на порядок по сравнению с «наивным» алгоритмом на типичных размерах входных данных. Применение предложенного алгоритма может значительно ускорить выполнение анализа представленности.

Ключевые слова
анализ представленности, экспрессия генов, кумулятивный алгоритм, эмпирическое распределение, корневая эвристика

Благодарности
Работа поддержана грантом Правительства Российской Федерации № 074-U01.

ALGORITHM FOR CUMULATIVE CALCULATION OF GENE SET ENRICHMENT STATISTIC

A.A. Sergushichev[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: alserg@rain.ifmo.ru

Article info
Received 08.07.16, accepted 10.08.16
doi: 10.17586/2226-1494-2016-16-5-956-959
Article in Russian
For citation: Sergushichev A.A. Algorithm for cumulative calculation of gene set enrichment statistic. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 5, pp. 956–959. doi: 10.17586/2226-1494-2016-16-5-956-959

Abstract
Methods for gene set enrichment analysis, widely-used for analysis of gene expression data, were studied. A problem of cumulative calculation of enrichment statistic was considered. For this problem an algorithm based on square root decomposition heuristic was developed. An asymptotic run-time complexity of the algorithm was found. Practical implementation showed an order of magnitude increase in performance compared to a naïve algorithm when run on typical input sizes. The developed algorithm can be used to improve significantly the performance of gene set enrichment analysis.

Keywords
gene set enrichment analysis, gene expression, cumulative algorithm, empirical distribution, square root decomposition
Aknowledgements
This work was supported by the Russian Federation Government Grant No. 074-U01.

Преранжированный анализ представленности наборов генов (pre-ranked Gene Set Enrichment Analysis) [1, 2] является широко распространенным методом для анализа экспрессии генов [3, 4]. Этот метод позволяет из заранее заданного списка функционально-связанных наборов генов выделить те из них, которые ведут себя неслучайным образом. Метод основан на вычислении статистики представленности, которая отражает степень координированности изменения экспрессии заданного набора генов. Для каж-

дого из входных наборов генов значение этой статистики сравнивается с ее эмпирическим фоновым распределением для случайных наборов генов такого же размера.

Во многих реализациях этого метода вычисление эмпирических распределений статистики представленности реализовано независимо для каждого набора генов [2, 5–7]. При таком подходе выполнение даже 1000 сэмплирований для каждого из 500 наборов генов требует не меньше 5 минут. С учетом же необходимой поправки на множественное сравнение желательно выполнение хотя бы 10000 сэмплирований, что очень долго.

Этот метод можно было бы ускорить за счет вычисления эмпирических фоновых распределений статистики представленности сразу для всех входных наборов. В таком варианте при наличии эффективного алгоритма кумулятивного вычисления статистики представленности можно из одной случайной перестановки генов получить значения одного элемента для всех фоновых распределений.

В настоящей работе предлагается такой эффективный алгоритм для кумулятивного вычисления статистики представленности.

Сначала введем формальное определение статистики представленности. Она представляет собой функцию $s_r(S, p)$, принимающую на вход упорядоченный по убыванию массив вещественных чисел S длиной N (весов генов) и список позиций в этом массиве p размера K (набор генов) и возвращающую вещественно число. Функция определяется через вспомогательный массив ES размера N+1:

$$ES_i = \begin{cases} 0, \text{ если } i = 0; \\ ES_i + \dfrac{1}{NS}|S_i|, \text{ если } 1 \le i \le N \text{ и } i \in p; \\ ES_i - \dfrac{1}{N-K}, \text{если } 1 \le i \le N \text{ и } i \notin p\ , \end{cases}$$

где NS является нормализующим фактором, равным $NS = \sum_{i \in p}|S_i|$.

Итоговым значением статистики является максимальный по модулю элемент ES: $s_r(S, p) = ES_{i^*}$, где $i^* = \arg\max |ES_i|$.

Кумулятивный вариант функции s_{rc} возвращает массив из значений s_r для всех префиксов p: $s_{rc}(S, p) = \{s_r(S, p[1..i]) | i \in 1..k\}$.

Значение статистики $s_r(S, p)$ удобно представлять в геометрическом виде. Для этого рассмотрим $N + 1$ точку (рисунок) с координатами нулевой точки (x_0, y_0) равными нулю $(0, 0)$. Для остальных N точек $1 \le i \le N$ координаты задаются по формулам:

$x_i = x_{i-1} + [i \notin p]$;

$y_i = y_{i-1} + [i \in p]|S_i|$.

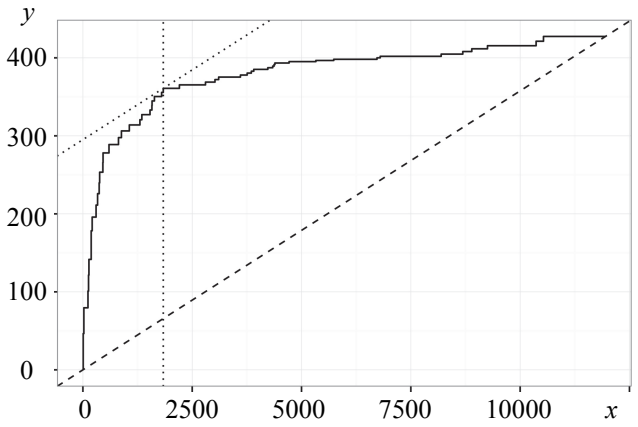

Рисунок. Графическое представление статистики представленности. Ось x соответствует позиции гена во входном массиве весов S. Ось y соответствует кумулятивной сумме модулей весов генов в наборе. Сплошная линия соответствует последовательности точек (x_0, y_0), (x_1, y_2), …, (x_N, y_N). Точка, соответствующая значению статистики для набора, является наиболее удаленной точкой графика от диагонали, проходящей через (x_0, y_0) и (x_N, y_N), выделенной штриховым пунктиром. Эта точка отмечена пересечением линий, выделенных точечным пунктиром: вертикальной линии и линии, параллельной диагонали

Несложно заметить, что $x_N = N - |p| = N - K$ и $y_N = \sum_{i \in p}|S_i| = NS$. Также промежуточные значения статистики ES_i могут быть вычислены как $ES_i = \frac{1}{NS} y_i - \frac{1}{N-K} x_i$, что пропорционально расстоянию от точки (x_i, y_i) до прямой, проходящей через (x_0, y_0) и (x_N, y_N).

Таким образом, определение итогового значения статистики s_r соответствует определению точки, максимально удаленной от диагонали. Возможность быстро обновлять такую точку при добавлении но-

вого гена в p позволила бы последовательно добавлять гены из p и тем самым эффективно находить кумулятивное значение статистики.

Рассмотрим, что происходит при добавлении нового гена с индексом g в набор p. В этом случае координаты точек (x_i, y_i) для $i < g$ не изменяются, а координаты всех точек (x_i, y_i) для $i \geq g$ изменяются на одинаковый вектор $(\delta x, \delta y) = (-1, |S_g|)$.

Применим корневую эвристику [8]: разобьем все $N+1$ точек на $b = \sqrt{N}$ одинаковых блоков размера b (для упрощения будем считать, что $N+1$ является точным квадратом b). Для каждого такого блока будем поддерживать в нем индекс наиболее удаленной точки от диагонали. Соответственно, наиболее удаленную точку среди всех $N+1$ точек можно найти за время $O(b)$, рассмотрев последовательно наиболее удаленные точки блоков.

Также будем в каждом блоке поддерживать выпуклую оболочку его точек. Заметим, что выпуклая оболочка для всех блоков, кроме содержащего точку g, остается неизменной. Для этого же блока перестроим оболочку за время $O(b)$. Это можно сделать за линейное время алгоритмом Грэхема [9], так как точки уже упорядочены по координате x. Таким образом, за время $O(b)$ мы можем поддерживать выпуклые оболочки для всех блоков.

Чтобы поддерживать индекс самой удаленной от диагонали точки внутри блока, воспользуемся знанием выпуклой оболочки. Самая удаленная от диагонали точка будет всегда лежать на выпуклой оболочке. При этом если выпуклая оболочка не изменяется, то, так как диагональ при добавлении гена вращается против часовой стрелки, индекс наиболее удаленной точки может только уменьшаться. Соответственно, для $b-1$ блоков, в которых не изменяется выпуклая оболочка, новый индекс самой удаленной точки можно найти, последовательно сравнивая текущую точку и точку на выпуклой оболочке, идущую сразу слева, и обновляя индекс до тех пор, пока следующая точка расположена ближе к диагонали. Для одного блока, в котором выпуклая оболочка перестраивается, наиболее удаленную точку найдем простым проходом по всем точкам выпуклой оболочки за не более чем линейное время $O(b)$. Так как индекс самой удаленной точки в блоке может сдвинуться вправо только при перестройке блока и на значение не больше b, то время, затраченное на уменьшение индексов, будет не больше $O(N+Kb)$.

Таким образом, предлагаемый алгоритм позволяет найти все промежуточные значения статистики s_r для набора генов размера K за время $O(Kb+N) = O\left(K\sqrt{N}+N\right)$. Для сравнения: алгоритм, вычисляющий значение статистики независимо для каждого префикса и эквивалентный независимому сэмплированию (наивный алгоритм), требует времени работы $O(K^2 \log K)$, так как вычисление для одного набора требует $O(K \log K)$ времени. При типичных значениях K=500 и N=20000 ускорение составляет примерно один порядок.

В работе предложен алгоритм кумулятивного вычисления статистики представленности для набора генов. Предложенный алгоритм работает на порядок быстрее наивной реализации. Это, в свою очередь, позволяет применить его в методе анализа представленности, что является задачей для дальнейшей работы.

Литература

1. Mootha V.K., Lindgren C.M., Eriksson K.-F. et al. PGC-1 α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes // Nature Genetics. 2003. V. 34. N 3. P. 267–273. doi: 10.1038/ng1180
2. Subramanian A., Tamayo P., Mootha V.K. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles // Proceedings of the National Academy of Sciences of the United States of America. 2005. V. 102. N 43. P. 15545–15550. doi: 10.1073/pnas.0506580102
3. Maciejewski H. Gene set analysis methods: statistical models and methodological differences // Briefings in Bioinformatics. 2014. V. 15. N 4. P. 504–518. doi: 10.1093/bib/bbt002
4. Tarca A.L., Bhatti G., Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity // PloS ONE. 2013. V. 8. N 11. P. e79217. doi: 10.1371/journal.pone.0079217
5. Yu G., Wang L.-G., Yan G.-R., He Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis // Bioinformatics. 2015. V. 31. N 4. P. 608–609. doi: 10.1093/bioinformatics/btu684
6. Väremo L., Nielsen J., Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods // Nucleic Acids Research. 2013. V. 41. N 8. P. 4378–4391. doi: 10.1093/nar/gkt111
7. Fang Z. GSEAPY: Gene Set Enrichment Analysis in Python

References

1. Mootha V.K., Lindgren C.M., Eriksson K.-F. et al. PGC-1 α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 2003, vol. 34, no. 3, pp. 267–273. doi: 10.1038/ng1180
2. Subramanian A., Tamayo P., Mootha V.K. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, vol. 102, no. 43, pp. 15545–15550. doi: 10.1073/pnas.0506580102
3. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics*, 2014, vol. 15, no. 4, pp. 504–518. doi: 10.1093/bib/bbt002
4. Tarca A.L., Bhatti G., Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS ONE*, 2013, vol. 8, no. 11, pp. e79217. doi: 10.1371/journal.pone.0079217
5. Yu G., Wang L.-G., Yan G.-R., He Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 2015, vol. 31, no. 4, pp. 608–609. doi: 10.1093/bioinformatics/btu684
6. Väremo L., Nielsen J., Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, 2013, vol. 41, no. 8, pp. 4378–4391. doi: 10.1093/nar/gkt111

[Электронный ресурс]. Режим доступа: https://github.com/BioNinja/gseapy, свободный. Яз. англ. (дата обращения 07.07.2016).
8. Иванов М. Sqrt-декомпозиция [Электронный ресурс]. Режим доступа: http://e-maxx.ru/algo/sqrt_decomposition, свободный. Яз. рус. (дата обращения 07.07.2016).
9. Кормен Т. и др. Алгоритмы: построение и анализ. 2-е изд. М.: Вильямс, 2005. 1296 с.

Автор

Сергушичев Алексей Александрович **–** аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, alserg@rain.ifmo.ru

7. Fang Z. *GSEAPY: Gene Set Enrichment Analysis in Python*. Available at: https://github.com/BioNinja/gseapy (accessed 07.07.2016).
8. Ivab\nov M. *Sqrt-dekompozitsiya* [Sqrt-decomposition]. Available at: http://e-maxx.ru/algo/sqrt_decomposition (accessed 07.07.2016).
9. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. *Introduction to Algorithms*. 2nd ed. Cambridge, MIT Press, 2006, 1312 p.

Author

Alexey A. Sergushichev **–** postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, alserg@rain.ifmo.ru