На правах рукописи

Сергушичев Алексей Александрович

Методы вычислительного анализа метаболических моделей для интерпретации транскриптомных и метаболомных данных

Специальность 05.13.18 — Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург — 2016

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Научный руководитель: **Артемов Максим**, PhD,
профессор-исследователь Университета ИТМО,
доцент Университета Вашнигтона в Сент-Луисе

Официальные оппоненты: **Макеев Всеволод Юрьевич**,
доктор физико-математических наук,
заведующий лабораторией,
Институт общей генетики РАН

Богачев Михаил Игоревич,
кандидат технических наук, доцент,
ведущий научный сотрудник,
Санкт-Птербургский государственный
электротехнический университет «ЛЭТИ»
имени В. И. Ульянова (Ленина)

Ведущая организация: федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный политехнический университет Петра Великого»

Защита состоится 22 декабря 2016 г. в 17 часов 00 минут на заседании диссертационного совета Д 212.227.06 при Санкт-Петербургском национальном исследовательском университете информационных технологий, механики и оптики по адресу: 197101, г. Санкт-Петербург, Кронверкский просп., д. 49., ауд. 431.

С диссертацией можно ознакомиться в библиотеке Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики по адресу: 197101, г. Санкт-Петербург, Кронверкский просп., д. 49. и на сайте:
http://fppo.ifmo.ru/?page1=16&page2=52&page_d=1&page_d2=146244.

Автореферат разослан «15» ноября 2016 г.

Ученый секретарь
диссертационного совета Д 212.227.06,
кандидат физико-математических наук,
доцент, Холодова С. Е.

Общая характеристика работы

Актуальность проблемы. С развитием технологий все большую роль в фундаментальных задачах биологии и медицины играет сбор больших объемов экспериментальных данных и последующий их анализ и интерпретация. Из-за больших объемов анализ и интерпретация вручную становятся практически невозможными, что влечет за собой необходимость разработки соответствующих вычислительных методов.

Одной из актуальных и быстро развивающихся областей биологии, требующих разработки новых методов для интерпретации данных, является изучение регуляции метаболизма (набора биохимических реакций, необходимых для жизнедеятельности клетки). Во-первых, стала ясна большая роль, которую метаболизм играет в биологических процессах, особенно в иммунной системе и раковых клетках. Во-вторых, появилась возможность широкого изучения метаболических процессов из-за удешевления технологий получения данных транскриптомного и метаболомного профилирования, отражающих активность ферментов и изменения в концентрациях веществ в клетке, соответственно.

Важным понятием с точки зрения интерпретации данных является метаболический путь – набор последовательных реакций, связанных одной функцией. Именно в терминах метаболических путей удобно интерпретировать данные, и поэтому можно рассматривать задачу интерпретации как задачу идентификации регулируемых метаболических путей и их взаимосвязей.

Исследовать метаболические пути удобно с помощью анализа метаболических моделей. В них обычно содержится информация о том, какие реакции могут происходить в клетке, как реакции связаны друг с другом, активность каких генов может регулировать реакции, какие из реакций относятся к стандартным метаболическим путям и т. д. С помощью таких моделей, а также экспериментальных данных, можно проанализировать, какие реакции являются наиболее важными, как они объединяются в метаболические пути и как эти пути взаимодействуют между собой.

Отметим, что хотя существуют наработки в области анализа метаболических путей, вопрос создания качественных эффективных вычислительных методов является актуальным. Даже для задачи, в которой требуется идентифицировать регулируемые метаболические пути среди набора заданных путей, имеется несколько конкурирующих подходов. Поиск же регулируемых путей без привязки к заранее заданному набору путей является еще более сложным. Для этого есть как минимум две причины. Во-первых, такие методы сильно зависят от рассматриваемой области и требуют доработки для каждой области отдельно. Во-вторых, многие такие задачи сводятся к оптимизационным задачам на графах, являющимся *NP*-трудными, что требует разработки специальных методов для нахождения оптимальных или субоптимальных решений.

Таким образом, рассматриваемая тема является актуальной, как с точки зрения развития вычислительных методов анализа метаболических моделей,

так и с точки зрения практической применимости для интерпретации экспериментальных данных в области изучения регуляции метаболизма.

В соответствии с паспортом специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ» диссертация относится к трем областям исследований: «3. Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий»; «6. Разработка новых математических методов и алгоритмов проверки адекватности математических моделей объектов на основе данных натурного эксперимента»; «7. Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели».

Целью работы является разработка и программная реализация набора эффективных вычислительных методов анализа метаболических моделей для идентификации регулируемых метаболических путей и их взаимосвязей по транскриптомным и метаболомным экспериментальным данным.

Основные задачи диссертационной работы состоят в следующем:

1. Разработка и реализация эффективного метода идентификации регулируемых путей в метаболических моделях на основе анализа представленности, без использования информации о связях между реакциями.
2. Разработка и реализация эффективного метода идентификации регулируемых путей и их взаимосвязей в метаболических моделях на основе подхода поиска активного модуля.
3. Разработка и реализация эффективного метода идентификации регулируемых путей и их взаимосвязей в метаболических моделях на основе подхода поиска активного модуля с использованием информации об атомной структуре метаболитов.

Научная новизна. В работе получены следующие новые научные результаты:

1. Разработан метод *FGSEA* (*Fast Gene Set Enrichment Analysis*) для проведения эффективного взвешенного анализа представленности функциональных наборов генов. Он позволяет идентифицировать регулируемые метаболические пути, используя только информацию из модели о множестве возможных реакций, их регуляции генами и их участии в метаболических путях. Метод является развитием существующего метода анализа представленности *GSEA* (*Gene Set Enrichment Analysis*). За счет использования разработанного алгоритма кумулятивного вычисления *GSEA*-статистики представленности, он позволяет достичь ускорения в сотни раз.
2. Разработан метод *GAM* (*Genes And Metabolites*) для выделения активных метаболических модулей с помощью анализа сети метаболических реакций. Он позволяет, используя информацию о связях реакций в метаболической модели, идентифицировать регулируемые метаболические пути и их взаимосвязи. По сравнению с существующими методами в

нем существует возможность использования нескольких вариантов представления сети реакций в виде графа в зависимости от входных данных. Также для возникающей в общем случае в методе обобщенной задачи поиска связного подграфа максимального веса (*GMWCS*, *Generalized Maximum Weight Connected Subgraph*), являющейся *NP*-трудной, разработан точный решатель.

3. Разработан метод *GATOM* (от *GAM* и *atom*) для выделения активных метаболических модулей с помощью анализа графа атомных переходов. Он позволяет идентифицировать регулируемые метаболические пути и их взаимосвязи, используя информацию о связях реакций в метаболической модели и о внутренней атомной структуре метаболитов. По сравнению с существующими методами в этом методе используется представление сети реакций в виде графа атомных переходов. Для учета структуры этого графа был сформулирован сигнальный вариант задачи *GMWCS* (*SGMWCS*, *Signal GMWCS*). Для задачи *SGMWCS*, также являющейся *NP*-трудной, разработан точный решатель.

Методы исследований. В работе используются методы дискретной математики, теории вероятности и математической статистики.

На защиту выносятся:

1. Метод *FGSEA* для проведения эффективного взвешенного анализа представленности функциональных наборов генов.
2. Метод *GAM* для выделения активных метаболических модулей с помощью анализа сети метаболических реакций.
3. Метод *GATOM* для выделения активных метаболических модулей с помощью анализа графа атомных переходов.

Достоверность научных положений, выводов и практических рекомендаций, полученных в диссертации, подтверждается корректным обоснованием постановок задач, точной формулировкой критериев, результатами экспериментов по использованию предложенных в диссертации методов и их анализом.

Теоретическое значение работы состоит в разработанном алгоритме кумулятивного подсчета *GSEA*-статистики представленности и асимптотической оценке времени его работы, формулировке задачи *SGMWCS*, формулировке задачи поиска активного модуля в виде задач *GMWCS* и *SGMWCS*, разработанных методах решения *NP*-трудных задач *GMWCS* и *SGMWCS*.

Практическое значение работы состоит в том, что разработанные методы могут и уже используются в настоящее время для изучения регуляции метаболизма, играющей особенно важную роль в работе иммунной системы млекопитающих и в развитии раковых опухолей.

Внедрение результатов работы. Программный пакет для анализа представленности, реализующий метод *FGSEA*, принят в библиотеку *R/Bioconductor* (http://bioconductor.org/packages/fgsea). Метод также внедрен в

рабочие процессы компании *Immuneering* (Кембридж, США). Метод *GAM* для анализа сетей реакций используется в компании *Elucidata* (Кембридж, США).

Апробация результатов работы. Основные результаты докладывались на следующих конференциях: 16th Workshop on Algorithms in Bioinformatics (WABI 2016), Оорхус, Дания; Всероссийской научной конференции по проблемам информатики «СПИСОК 2016», СПбГУ, Матмех; Moscow Conference on Computational Molecular Biology (MCCMB'15), 2015, Москва; Cold Spring Harbor Laboratory meeting on Systems Biology: Networks, 2015, Колд-Спринг-Харбор, США; IV международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине», 2014, Казань; Metabolism and Immunity: A Rediscovered Frontier, 2014, Дублин, Ирландия.

Личный вклад автора. Решение задач диссертации, разработанные методы *FGSEA*, *GAM* и *GATOM* принадлежат лично автору, а разработка решателей для задач *GMWCS* и *SGMWCS* была выполнена в соавторстве с А. А. Лободой.

Публикации. Основные результаты по теме диссертации изложены в девяти публикациях, в том числе одна из них в российском журнале из списка рекомендованных ВАК, и шесть, входящих в базы *Scopus* и *Web of Science*.

Регистрация программ. Автором по теме диссертации было получено свидетельство о регистрации программы для ЭВМ: Сергушичев А. А. Программа для быстрого анализа представленности метаболических путей по упорядоченному списку генов с весами // Свидетельство №2016 660664 от 20.09.2016.

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и одного приложения. Объем диссертации – 126 страниц с 40 рисунками и двумя таблицами. Список литературы содержит 101 наименование.

Содержание работы

В **первой главе** приводится обзор работ, посвященных анализу транскриптомных и метаболомных данных, методам анализа метаболических моделей, формальным постановкам задачи поиска активного модуля и подходам к их решению.

Сначала вводятся основные понятия предметной области. Рассматривается понятие биохимической реакции. Описывается связь реакций, ферментов, генов и метаболитов. Вводится понятие метаболического пути. Приводится информация о способах получения транскриптомных и метаболомных данных и их особенностях с точки зрения анализа. Описывается метод анализа дифференциальной экспрессии, являющийся базовым методом анализа такого рода данных.

Затем рассматриваются полногеномные *метаболические модели клетки*. Такие модели обязательно включают в себя список всех возможных реакций в

клетке. Кроме этого могут быть указаны: 1) отделы клетки, в которых протекают реакции; 2) набор правил регуляции реакций; 3) кинетические уравнения для скорости реакций. 4) термодинамические ограничения, задающие возможные направления реакций; 5) параметры связи с окружающей средой, такие как наличие веществ во внешней среде; 6) дополнительная аннотация, такая как принадлежность реакции тому или иному метаболическому пути.

В зависимости от степени детализации моделей для их анализа могут применяться разные методы. Самым простым подходом является анализ представленности. В нем используется только информация о том, как реакции группируются в метаболические пути и какими генами они регулируются. Анализ представленности состоит в том, чтобы выделить метаболические пути, которые хорошо представлены среди генов с сильной индивидуальной регуляцией.

Другим подходом является поиск «активного модуля». В нем используется информация о том, как реакции связаны друг с другом. Целью такого подхода является выделение связного набора реакций, имеющих регулярное поведение в экспериментальных данных – активного модуля. Такой подход позволяют найти связи между известными метаболическими путями, а также выделить ранее не описанные.

Кроме этого, выделяется группа методов, которые рассматривают атомную структуру метаболитов. Некоторые из таких методов используют эту информацию для поиска потенциальных метаболических путей.

Затем рассматриваются методы анализа представленности, в частности, метод *GSEA*. После рассматриваются варианты постановки задачи поиска активного модуля. Особое внимание уделяется задаче поиска подграфа максимального веса (*Maximum Weight Connected Subgraph*, *MWCS*) и ее вариантам.

На основе результатов обзора формулируются цель и решаемые задачи диссертации.

Вторая глава посвящена решению задачи разработки эффективного вычислительного метода идентификации регулируемых метаболических путей с помощью анализа представленности. В этой главе предлагается метод *FGSEA*, развивающий существующий ранее метод *GSEA*. Разработанный метод состоит в том, чтобы вместо независимого подсчета фонового распределения для каждого данного размера набора генов строить эти распределения одновременно. Для этого был разработан алгоритм кумулятивного вычисления *GSEA*-статистики. Применение этого алгоритма позволило достичь ускорения на два порядка по сравнению с методом *GSEA*. Это, в свою очередь, позволяет увеличить число рассматриваемых случайных наборов для построения эмпирического фонового распределения и увеличить точность метода.

Сначала идея метода рассматривается для статистики представленности s_m среднего веса генов в наборе p, определяемой по формуле:

$$s_m(p) = \frac{1}{|p|}\Sigma_{i \in p} S_i,$$

где S_i – вес i-го гена.

Идея метода состоит в том, чтобы переиспользовать результаты вычисления эмпирического фонового распределения для разных наборов генов. Такой подход является корректным, так как для оценки фонового распределения случайные наборы должны быть независимы только для каждого входного набора в отдельности.

Таким образом, вместо генерации nm независимых случайных наборов можно генерировать только n случайных наборов размера K, где n – размер выборки для оценки фонового распределения, m – число рассматриваемых наборов генов, а K – максимальный размер набора генов. Из одного i-го случайного набора π_i размера K можно получить случайные наборы для всех размеров $k \leqslant K$, взяв его префиксы: $\pi_{i,k} = \pi_i[1..k]$.

Следующим шагом является вычисление значений статистики представленности для всех наборов $\pi_{i,j}$. Вместо того, чтобы вычислять статистику представленности независимо для всех наборов, можно вычислять значения одновременно для всех $\pi_{i,j}$ для фиксированного i. Для статистики s_m достаточно за один проход кумулятивно вычислить частичные суммы рангов генов и поэлементно поделить их на длины соответствующих префиксов:

$$s_m(\pi_i[1..k]) = \frac{1}{k}\Sigma_{i\in\pi_i[1..k]}S_i.$$

Идею использования кумулятивного вычисления статистики для ускорения вычисления *P*-значений можно применить и для *GSEA*-статистики. Однако, для случая *GSEA*-статистики кумулятивное вычисления требует разработки отдельного алгоритма. Для простоты, сначала рассматривается только положительная мода *GSEA*-статистики s_r^+. Она вычисляется по формуле $s_r^+(p) = \mathrm{ES}_{i^+}$, где $i^+ = \arg\max_i \mathrm{ES}_i$, а

$$\mathrm{ES}_i = \begin{cases} 0, & \text{если } i = 0, \\ \mathrm{ES}_{i-1} + \frac{1}{\mathrm{NS}}|S_i|, & \text{если } 1 \leqslant i \leqslant N \text{ и } i \in p, \\ \mathrm{ES}_{i-1} - \frac{1}{N-k}, & \text{если } 1 \leqslant i \leqslant N \text{ и } i \notin p. \end{cases}$$

В алгоритме кумулятивного вычисления *GSEA*-статистики для набора p размера K она рассматривается с геометрической точки зрения. Для этого вводится $N + 1$ точка (рисунок 1) с координатами (x_i, y_i) для $0 \leqslant i \leqslant N$, определяемых как:

$$(x_0, y_0) = (0, 0), \tag{1}$$

$$x_i = x_{i-1} + [i \notin p], \qquad \forall i \in 1..N, \tag{2}$$

$$y_i = y_{i-1} + [i \in P] \cdot |S_i|, \qquad \forall i \in 1..N. \tag{3}$$

Значение *GSEA*-статистики можно вычислить, зная положение самой удаленной от диагонали $\big((x_0, y_0), (x_N, y_N)\big)$ точки среди (x_i, y_i). Предложенный

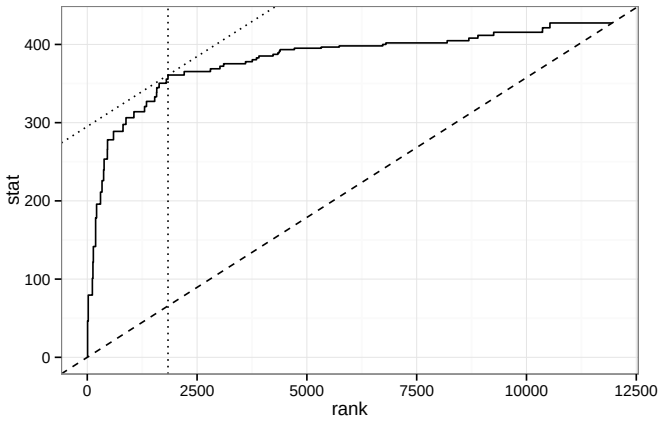

Рисунок 1 – График, соответствующий геометрическому представлению *GSEA*-статистики. Значение статистики представленности соответствует точке на графике (пересечение линий с мелким пунктиром), наиболее удаленной от диагонали (линия с крупным пунктиром)

алгоритм позволяет за время $O(\sqrt{K})$ обновить положение этой точки после добавления очередного гена. В алгоритме используется корневая эвристика, состоящая в том, что все точки можно разделить на $\sqrt{K}$ блоков. В каждом из них можно эффективно поддерживать выпуклую оболочку и обновлять положение самой удаленной точки среди точек блока. Действительно, можно заметить, что при добавлении гена только в одном блоке происходят серьезные изменения, а все остальные блоки либо остаются неизменными, либо сдвигаются на одинаковый вектор. Это позволяет полностью обновить за время $O(\sqrt{K})$ сильно изменившийся блок и за амортизированное время $O(\sqrt{K})$ обновить все остальные блоки.

Суммарно, предложенный алгоритм позволяет кумулятивно вычислить значения *GSEA*-статистики представленности $s_r(\pi[1..k])$ за время $O(K\sqrt{K})$. Простая же реализация с независимым вычислением статистики требует $O(K^2 \log K)$ операций. Таким образом, производительность увеличивается в $O(K\frac{\log K}{\sqrt{K}})$ раз.

Далее приведены результаты экспериментального анализа алгоритма кумулятивного вычисления *GSEA*-статистики, подтверждающие оценку $O(K\sqrt{K})$ на время работы алгоритма.

Затем разработанный метод сравнивается с существующей референсной реализацией метода *GSEA*. Оба метода были запущены с генераций 1000 и 10000 случайных наборов для оценки фонового распределения. Время работы референсной реализации составило 96 с для 1000 случайных наборов и 1048 с – для 10000. Для реализации *FGSEA* время работы составило $0{,}8$ с для 1000 случайных наборов и $5{,}5$ с – для 10000. Таким образом, для 1000 случайных наборов ускорение составило 120 раз, а для 10000 – 190 раз. Полученные *P*-значения достаточно хорошо согласуются для обоих методов (коэффициент корреляции $99{,}98\%$, рисунок 2), неполное соответствие значений объясняется недетерминированностью процесса.

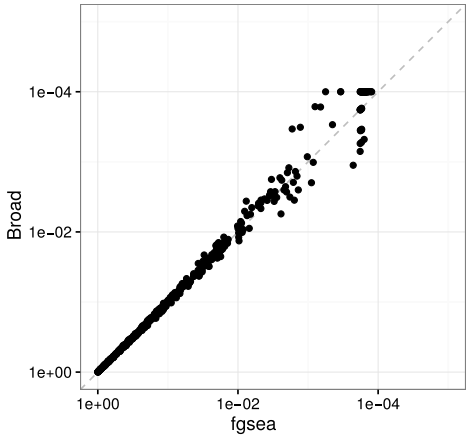

Рисунок 2 – Номинальные *P*-значения, вычисленные двумя методами. По оси абсцисс – значения для метода *FGSEA*, по оси ординат – для референсной реализации

Кроме того было показано, что начиная с 10000 случайных наборов становится возможно использовать стандартные методы поправки на множественное сравнение, например, метод Бенджамини-Хохберга. Этот метод лучше контролирует уровень ошибок, чем метод, разработанный специально для референсной реализации *GSEA*. Поэтому использование метода *FGSEA*, позволяет получать более точные результаты по сравнению с методом *GSEA*.

В конце приводится пример применения метода для определения важных метаболических путей при активации Т-клеток.

В **третьей главе** предлагается метод *GAM* для поиска активных модулей в метаболических моделях, основанный на сведении к задаче поиска связного подграфа максимального веса. Предложенный метод развивает подход, предложенный М. Диттрихом и соавторами для сетей белок-белковых взаимодействий.

Сначала описывается общая схема предлагаемого метода, представленная на рисунке 3. Первым шагом из базы данных *KEGG* выделяются реакции, потенциально возможные в выбранном организме (рисунок 3A). Затем при наличии транскриптомных данных удаляются реакции без экспрессированных генов (рисунок 3B). Далее набор реакций представляется в виде графа (рисунок 3C). На основе результатов анализа дифференциальной экспрессии вершинам и ребрам графа приписываются веса: положительные – значимо регулируемым генам и метаболитам, отрицательные – не значимым (рисунок 3D). В полученном графе выделяется активный модуль с помощью решения задачи *GMWCS* (рисунок 3E). Дополнительные процедуры постобработки приводят к финальному модулю (рисунок 3F).

При построении сети выполняется несколько шагов предобработки. Во-первых, чтобы избежать дублирования, удаляются отдельный шаги многоступенчатых реакций и сохраняется только соответствующая общая реакция. Во-вторых, удаляются неспецифичные метаболиты, создающие чрезмерную связность, такие как, например, простые неорганические вещества.

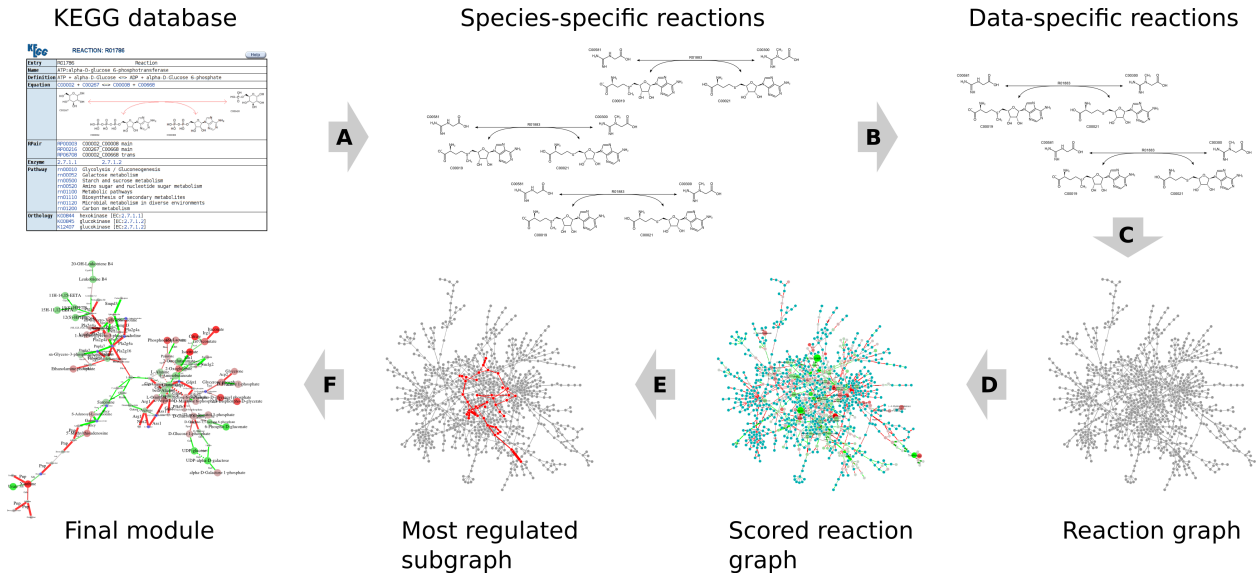

Рисунок 3 – Схема предлагаемого метода *GAM*

Затем рассматривается вопрос представления сети реакций в виде графа, что необходимо для использования сведения к задаче *GMWCS*. Из-за существования мультимолекулярных реакций, в которых участвует более одного вещества с одной из сторон, невозможно тривиальным образом представить набор возможных реакций в виде простого графа. Рассматриваются несколько вариантов возможного представления, принципиально разделяемых на два класса: в одном и реакции, и вещества отображаются на вершины, в другом, вещества отображаются на вершины, а реакции на ребра. В случае наличия только транскриптомных данных рекомендуется выбирать вершинное представление реакций. При наличии же метаболомных данных рекомендуется выбирать вершинно-реберное представление.

Схема назначения весов генов и метаболитов адаптирована из работы М. Диттриха. Кратко, исходя из предположения, что P-значения подчиняются смеси бета- и равномерного распределений, находятся параметры этой смеси для генов и метаболитов, присутствующих в сети. Вес вычисляется по формуле: $w(x) = (\alpha - 1)(\log x - \log \tau)$, где x – P-значение из теста дифференциальной экспрессии, α – найденный параметр бета-распределения и τ – порог, позволяющий изменять число генов и метаболитов с положительным весом.

После назначения весов возникает экземпляр задачи *GMWCS*. В некоторых случаях можно рассматривать задачу *SMWCS* (*Simple Maximum Weight Connected Subgraph Problem*), отличающуюся от задачи *GMWCS* отсутствием весов ребер.

Дальше рассматривается разработанный решатель для задачи *GMWCS*. Он основан на сведении к задаче целочисленного линейного программирования и последующего ее решения с помощью библиотеки *IBM ILOG CPLEX*. Решатель включает в себя несколько компонент:

1. Правила предобработки, позволяющие упростить задачу.

2. Правила декомпозиции, позволяющие разбить задачу на более мелкие, из решений которых можно восстановить решение исходной задачи.
3. Формулировка ограничений на связность подграфа в виде набора линейных ограничений.

После этого приводится описание разработанного веб-сервиса, реализующего предлагаемый метод и доступного по адресу http://genome.ifmo.ru/shiny/gam. Этот сервис позволил собрать набор реальных данных для исследования работы метода.

Главу завершают результаты экспериментального исследования. Приводится описание результатов анализа работы метода на искусственно сгенерированных данных. Для одного эксперимента случайным образом выбирался набор метаболических путей, которые составляли «правильный» активный модуль. Затем для генов модуля генерировали *P*-значения из бета-распределения $B(\alpha, 1)$, а для остальных генов – из равномерного распределения. Параметр α выбирался из равномерного распределения $U(0{,}01, 0{,}5)$, что соответствует распределению α в собранном наборе реальных данных. Было проведено сравнение точности восстановления отдельных генов с базовым методом, выбирающим гены с самыми значимыми *P*-значениями. Было показано, что метод *GAM* работает лучше базового метода в более сложных случаях, когда значение α больше $0{,}1$. Кроме этого была рассмотрена точность идентификации регулируемых метаболических путей с помощью применения метода Фишера к набору генов найденного модуля. Было показано, что, во-первых, такой метод контролирует уровень ошибок первого рода, а, во-вторых, позволяет идентифицировать больше регулируемых метаболических путей по сравнению с комбинацией метода Фишера и выбора генов с наиболее значимыми *P*-значениями.

Затем была проанализирована работа на реальных транскриптомных и метаболомных данных. Так как для этих данных сложно оценить точность метода, проводился анализ числа идентифицированных метаболических путей. Было показано, что метод *GAM* позволяет обнаружить больше путей, чем соответствующий базовый метод.

В конце приводится пример работы модуля на данных активации макрофагов (рисунок 4). С помощью анализа была выявлена важность нескольких метаболических путей. Их регуляция подтверждается литературными источниками.

Таким образом, разработанный метод *GAM*, во-первых, за счет использования структуры графа в сложных ситуациях позволяет лучше определять принадлежность индивидуальных генов активному модулю, чем метод, ориентирующийся только на *P*-значения. Во-вторых, метод может использоваться для идентификации регулируемых метаболических путей как среди заданного списка, так и идентификации неизвестных путей, так как информация об аннотации реакций метаболическими путями явным образом не используется.

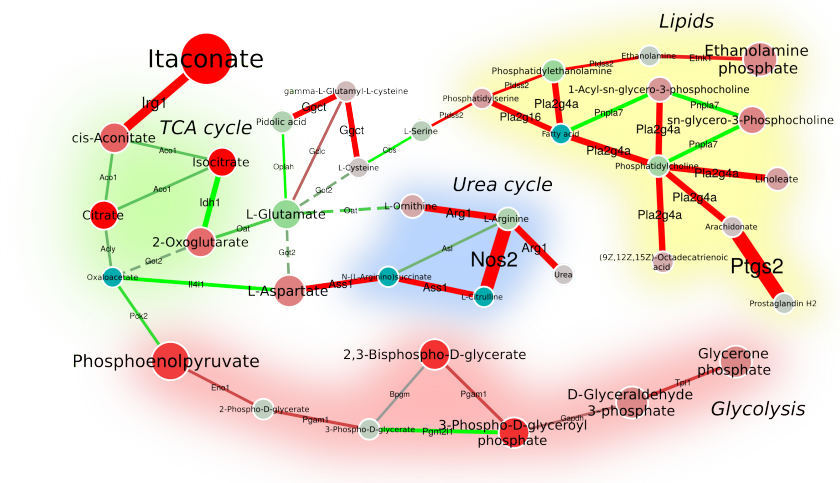

Рисунок 4 – Модуль, найденный с помощью метода *GAM*, при сравнении неактивированных и классически активированных макрофагов. Красным цветом обозначены метаболиты и гены, концентрация и экспрессия которых уменьшается при активации, зеленым – для которых увеличивается

В **четвертой главе** предлагается метод *GATOM* для поиска активных модулей в метаболических моделях, основанный на использовании графа атомных переходов. Метод является развитием метода *GAM*, но за счет работы на атомном уровне позволяет более качественно идентифицировать регулируемые метаболические пути.

Сначала рассматривается используемый граф атомных переходов. В этом графе вершинами являются отдельные углеродные атомы метаболитов, а ребро соединяют пару атомов, если существует реакция, переводящая их друг в друга. Приводится пример реакции *L-Cysteine* + *Glutathione* + $NADP^+$ <=> *S-Glutathionyl-L-cysteine* + *NADPH*, которая позволяет последовательно соединить метаболиты *L-Cysteine* и *Glutathione* через *S-Glutathionyl-L-cysteine* в методе *GAM*. Такое соединение отличается от последовательного превращения одного вещества в другое через промежуточное вещество, что затрудняет интерпретацию. Использование графа атомных переходов позволяет избежать таких соединений.

Граф атомных переходов строится на основе базы *KEGG RPAIR*. В этой базе для большого числа реакций (примерно 8000) указаны переходы атомов субстратов в атомы продуктов. Из-за того, что не во всех вхождениях одного метаболита в реакции в записях совпадает нумерация, выполняется дополнительная ее нормализация.

Показывается, что из-за «расслоения» сети реакций, получающийся граф имеет сложную повторяющуюся структуру. Она делает невозможным прямое сведение задачи к задаче *GMWCS*, так как при этом существовали бы систематические искажения в оценке модулей, зависящие от близости разных атомов одного вещества в рассматриваемом графе.

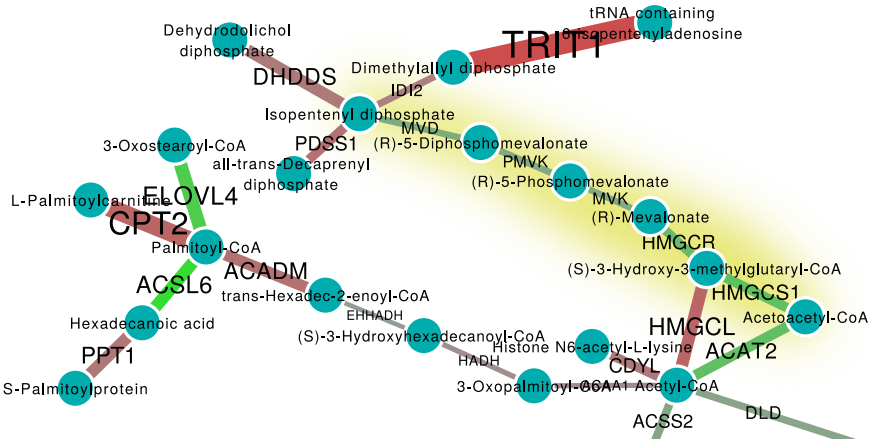

Рисунок 5 – Фрагмент активного метаболического модуля при сравнении образцов глиомы с мутацией в гене TP53 и без нее

Чтобы учесть структуру графа вводится сигнальный вариант задачи *GWMCS – SGMWC*. В нем вместо обычных весов каждой вершине и ребру ставится в соответствие сигнал. Каждому сигналу присваивается вес, причем сигналу с отрицательным весом может соответствовать либо одна вершина, либо одно ребро. Соответственно, для каждого гена (метаболита), для которого в методе *GAM* назначался бы положительный вес, вводится один сигнал на все его вхождения. Для генов (метаболитов) с отрицательным весом вводится по отдельному сигналу для каждого вхождения. Вес модуля определяется как сумма всех весов его сигналов без учета повторов – каждый сигнал учитывается в весе не больше одного раза. При такой схеме перестает быть выгодным добавлять несколько раз один и тот же ген (метаболит) с положительным весом.

На основе решателя для задачи *GMWCS*, разработанного для метода *GAM*, был разработан решатель для задачи *SGMWCS*. Из-за более общего вида задачи были модифицированы правила предобработки и разработана другая схема декомпозиции. В новой схеме задача разбивается на последовательность задач с фиксированным корнем, которые решаются параллельно.

Затем описываются проведенные экспериментальные исследования. Схема экспериментов повторяет схему из предыдущей главы. Сначала на сгенерированных данных было показано, что, так же, как и метод *GAM*, метод *GATOM* позволяет лучше определять отдельные гены из активного модуля по сравнению с базовым методом, использующим только *P*-значения, при $\alpha > 0{,}1$. Было показано, что в получающемся активном модуле можно находить известные метаболические пути с помощью метода Фишера, причем такой метод работает лучше применения метода Фишера к самым значимым генам. Аналогично, на реальных данных было показано, что метод *GATOM* позволяет находить значительно больше метаболических путей, чем базовый метод и метод *GAM*.

В конце главы приводится пример работы метода *GATOM* на транскриптомных данных в раковых образцах глиомы. Проводилось сравнение образцов с мутацией в гене *TP53* и без мутации. Фрагмент получившегося модуля пред-

ставлен на рисунке 5. В нем хорошо заметен метаболический путь биосинтеза мевалоната (выделен цветом). Регуляция этого пути подтверждается статьей Лаецца и соавторов, в которой показывается, что именно мутация в гене *TP53* влечет такие изменения. Важно отметить, что этот фрагмент появляется в модуле из-за сильного изменения в экспрессии гена *TRIT1*, субстрат которого может получаться только через этот метаболический путь. В то же время при использовании обычного графа метаболических реакций, находятся другие, опосредованные, способы соединить этот ген с другими компонентами модуля, не включающие метаболический путь биосинтеза мевалоната.

Таким образом, разработанный метод *GATOM*, так же, как и метод *GAM*, позволяет идентифицировать регулируемые метаболические пути, не используя в явном виде информацию об известных метаболических путях. Соответственно, он может быть использован как для идентификации существующих регулируемых метаболических путей и их взаимосвязей, так и для поиска новых путей. По сравнение с методом *GAM*, метод *GATOM* позволяет находить более тонкие эффекты за счет использования графа атомных переходов.

Заключение

В диссертационном исследовании получены следующие основные результаты:

1. Разработан метод быстрого взвешенного анализа представленности наборов генов. Он основан на алгоритме для кумулятивного подсчета значения статистики представленности. Метод реализован в виде *R*-пакета *fgsea* и доступен в открытый библиотеке *Bioconductor* (http://bioconductor.org/packages/fgsea).
2. Разработан метод поиска активного метаболического модуля с помощью анализа сети метаболических реакций. Он состоит в представлении сети реакций и исходных данных в виде взвешенного графа и последующего решения задачи поиска связного подграфа максимального веса. Метод доступен для использования в виде веб-сервиса *Shiny GAM* (http://genome.ifmo.ru/shiny/gam/). Для решения обобщенного варианта задачи, возникающего при одновременном наличии данных транскриптомного и метаболомного профилирования, был разработан открытый точный решатель (https://github.com/ctlab/gmwcs-solver/).
3. Разработан метод поиска активного метаболического модуля с помощью анализа графа атомных переходов. Использование графа атомных переходов позволяет избежать некоторых типов соединений и лучше соответствует структуре метаболических путей. Метод также доступен в веб-сервисе *Shiny GAM*. Для учета повторяющейся структуры графа атомных переходов была разработан сигнальный вариант обобщенной задачи поиска связного подграфа максимального веса. Для этого варианта также был разработан открытый точный решатель (https://github.com/ctlab/sgmwcs-solver/).

Статьи в журналах из перечня ВАК

1. ***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. — 0,2 п. л.

Публикации в рецензируемых изданиях, индексируемых Web of Science или Scopus

2. *Loboda A. A.*, *Artyomov M. N.*, ***Sergushichev A. A.*** Solving generalized maximum-weight connected subgraph problem for network enrichment analysis // Algorithms in Bioinformatics: 16th International Workshop, WABI 2016, Proceedings. — Springer Int. Pub., 2016. — Pp. 210–221. — 0,5 п. л. / 0,3 п. л.
3. *Izreig S.*, *Samborska B.*, *Johnson R. M.*, ***Sergushichev A.***, [et al.] The miR-17-92 microRNA cluster is a global regulator of tumor metabolism // Cell Rep. — 2016. — Vol. 16, no. 7. — Pp. 1915–1928. — 0,9 п. л. / 0,1 п. л.
4. ***Sergushichev A. A.***, *Loboda A. A.*, *Jha A. K.*, *Vincent E. E.*, [et al.] GAM: a web-service for integrated transcriptional and metabolic network analysis // Nucleic Acids Research. — 2016. — Vol. 44, W1. — W194–W200. — 0,65 п. л. / 0,6 п. л.
5. *Lampropoulou V.*, ***Sergushichev A.***, *Bambouskova M.*, *Nair S.*, [et al.] Itaconate links inhibition of succinate dehydrogenase with macrophage metabolic remodeling and regulation of inflammation // Cell Metab. — 2016. — Vol. 24, no. 1. — Pp. 158–166. — 0,7 п. л. / 0,25 п. л.
6. *Vincent E. E.*, ***Sergushichev A. A.***, *Griss T.*, *Gingras M.*, [et al.] Mitochondrial phosphoenolpyruvate carboxykinase regulates metabolic adaptation and enables glucose-independent tumor growth // Molecular Cell. — 2015. — Vol. 60, no. 2. — Pp. 195–207. — 1,3 п. л. / 0,5 п. л.
7. *Jha A. K.*, *Huang S. C.*, ***Sergushichev* A. A.**, *Lampropoulou V.*, [et al.] Network integration of parallel metabolic and transcriptional data reveals metabolic modules that regulate macrophage polarization // Immunity. — 2015. — Vol. 42, no. 3. — Pp. 419–430. — 1,3 п. л. / 0,3 п. л.

Другие публикации

8. ***Sergushichev A.*** An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation // bioRxiv. — 2016. — DOI: 10.1101/060012. — URL: http://biorxiv.org/content/early/2016/06/20/060012. — 0,4 п. л.
9. ***Сергушичев А. А.*** Алгоритм для быстрого анализа перепредставленности генов // Всероссийская научная конференция по проблемам информатики *СПИСОК*. — СПб. : ВВМ. СПбГУ, 2016. — С. 517–524. — 0,25 п. л.

Тиражирование и брошюровка выполнены в учреждении «Университетские телекоммуникации».
197110, Санкт-Петербург, Саблинская ул., 14. Тел. (812) 2334669. Объем 1,0 у.п.л. Тираж 100 экз.