Санкт-Петербургский государственный университет

ВЯТКИН Никита Сергеевич

Выпускная квалификационная работа

Вариативная идентификация природных соединений по масс-спектрам

Уровень образования: бакалавриат
Направление 01.03.01 «Математика»
Основная образовательная программа СВ.5000.2019 «Математика»

Научный руководитель:
профессор, д. ф.-м. н.
Степанов Алексей Владимирович

Рецензент:
младший профессор, Факультет компьютерных наук,
Саарский университет, к. ф.-м. н.
Гуревич Алексей Александрович

Санкт-Петербург
2023 г.

Содержание

Введение

Природные соединения (англ. natural products) — органические соединения, синтезируемые микроорганизмами. Они не являются необходимыми для выживания, но дают организмам, которые их производят, эволюционное преимущество. Пептидные природные соединения (ППС) представляют собой фармакологически важный класс природных соединений; многие его представители обладают антимикробными, противораковыми и противовирусными свойствами [1]. ППС состоят из аминокислот, соединённых пептидными связями, но в отличии от обычных пептидов и белков, ППС может содержать нестандартные аминокислоты (более 100 видов), редкие пост-трансляционные модификации и иметь сложную топологию, например, циклическую или разветвлённую структуру. Эти особенности существенно затрудняют поиск новых ППС, несмотря на их высокую ценность с точки зрения медицины.

Наиболее распространённый метод для идентификации ППС в природных образцах — тандемная масс-спектрометрия. Молекулы данного образца ионизируются, и первый масс-спектрометр разделяет эти ионы по их отношению массы к заряду (часто обозначается как m/z). Ионы с определённым соотношением m/z отбираются и поступают во второй масс-спектрометр. Он расщепляет эти ионы на более мелкие фрагменты, разделяет по соотношению m/z и производит подсчёт доли каждого из фрагментов. Так получается *масс-спектр* (*тандемный масс-спектр*, *спектр*) вещества.

В то время, как миллионы тандемных масс-спектров пептидных природных соединений были получены и размещены в базе Global Natural Products Social (GNPS) [2], для подавляющего большинства из них до сих пор неизвестно вещество, породившее спектр. Поэтому возникает задача *идентификации* — нахождения в базе данных веществ того ППС, которое сгенерировало определённый масс-спектр. Из-за особенностей строения пептидных природных соединений идентификация их спектров значительно труднее традиционной идентификации белков и пептидов в протеомике. В частности, классические инструменты из протеомики не применимы в этой задаче.

Нестандартное строение лишь одна из двух основных трудностей идентификации пептидных природных соединений. Во многих случаях вещество, породившее спектр, отсутствует в базе данных, тогда как, например, с одной изменённой или отсутствующей аминокислотой – присутствует. Задача идентификации по масс-спектру неизвестного ППС из его известных вариантов называется *вариативной идентификацией*, в отличие от стандартной идентификации, когда ППС присутствует в базе данных.

Поскольку большинство ППС образуют семейства близких по строению соединений, вариативная идентификация имеет решающее значение для открытия новых пептидных природных соединений. Поиск вариантов известных ППС важен, так как они иногда более эффективны с клинической точки зрения, чем другие представители семейства.

Для идентификации пептидных природных соединений создано несколько алгоритмов, но качество их работы ещё далеко от идеала. Примерами таких алгоритмов являются Dereplicator

[3] и VarQuest [4]. В то время, как Dereplicator разработан для стандартной идентификации, VarQuest способен производить вариативную идентификацию пептидных природных соединений. Для данного масс-спектра VarQuest проводит поиск в базе данных и выдаёт список из возможных кандидатов — известных ППС, с указанием аминокислоты, модификация которой может привести к ППС, породившему исходный спектр. При этом алгоритм устроен так, что способен обнаруживать только пептидные природные соединения, отличающиеся одной модификацией. Однако разнообразие ППС и их модификаций столь велико, что для многих полученных масс-спектров ещё не известно ни соединение, породившее спектр, ни один из его вариантов с одной модификацией. С другой стороны, хоть модификации и бывают весьма разнообразными, некоторые из них встречаются в природе намного чаще, чем другие. В связи с этим возникает потребность расширить функциональность современных алгоритмов для вариативной идентификации ППС.

Целью данной работы является реализация алгоритма вариативной идентификации масс-спектров пептидных природных соединений с учётом двух возможных модификаций. При этом одна модификация предполагается распространённой (из небольшого списка конкретных высокочастотных модификаций), а вторая — произвольная, как это было изначально в алгоритме VarQuest.

1. Постановка задачи

1.1. Терминология

Мы будем называть *ori-ППС* — пептидное природное соединение, породившее данный масс-спектр S.

При работе с пептидными природными соединениями используется термин *обобщённые пептидные связи* [3]. Помимо классических пептидных связей он включает в себя эфирные C-O связи, а также C-C связи между тиазолами/оксазолами, дегидроаланинами/дегидробутиринами и другими аминокислотами. Поэтому при разделении ППС на части обобщёнными пептидными связями, могут получаться очень нестандартные фрагменты, которые мы будем для простоты называть аминокислотами.

1.2. ППС-графы и их модификации

В рамках данной работы я рассматриваю пептидные природные соединения как ориентированные графы. Пусть P — некоторое пептидное природное соединение.

Определение 1. *Граф пептидного природного соединения* (*ППС-граф*) P — это связный ориентированный граф $G = G(P) = (V, E, M)$, где V – множество вершин (аминокислоты P), E – множество рёбер (пептидные связи), а $M : V \to \mathbb{R}$ — массы аминокислот в дальтонах (Da). Петли и кратные рёбра допускаются, направление ребра выбирается от аминогруппы к карбоксильной группе.

Такая формализация структуры пептидных природных соединений является общепринятой для задач (вариативной) идентификации, поскольку при фрагментации ионов в масс-спектрометре пептидные связи более остальных подвержены разрыву. Это означает, что получающиеся фрагменты в основном будут состоять из целых аминокислот.

Определение 2. Два ППС-графа $G_1 = (V_1, E_1, M_1)$ и $G_2 = (V_2, E_2, M_2)$ будем называть ε-*изоморфными*, если существует биекция $\sigma : V_1 \to V_2$, которая устанавливает изоморфизм G_1 и G_2 как обычных ориентированных графов, а также $\forall\, v \in V_1 : |M_1(v) - M_2(\sigma(v))| < \varepsilon$.

В дальнейшем ε всегда будет равно 0.02, а под изоморфностью ППС-графов я буду понимать 0.02-изоморфность. Эта погрешность в равенстве масс соответствующих аминокислот необходима, поскольку мы работаем с данными реальных физических приборов (масс-спектрометров), которые не могут определять характеристики молекул, а именно отношение m/z ионов, с идеальной точностью.

Определение 3. *Модификацией* при вершине v ППС-графа называется каждая из следующих операций:

1. *замена* — изменение $M(v)$ (массы вершины v);

2. *удаление* вершины v и соответствующая корректировка масс соседних вершин (см. далее);

3. *вставка* новой вершины, инцидентной v, и соответствующая корректировка массы вершины v;

Удаление вершины v не должно нарушать связность и структуру ППС-графа, поэтому я допускаю удаление только тех вершин, входящая ($indeg(v)$) и исходящая ($outdeg(v)$) степени которых не превышают единицы. Если $indeg(v) = outdeg(v) = 1$ и $(u, v), (v, w) \in E$, то после удаления v необходимо добавить ребро (u, w) между соседями вершины v. Таким образом, их входящая и исходящая степени останутся прежними. Если одно из чисел $indeg(v), outdeg(v)$ равно единице, а второе — нулю, то при удалении v масса соседней с ней вершины должна быть увеличена на массу H (≈ 1 Da) или OH (≈ 17 Da) при $indeg(v) = 1$ или $outdeg(v) = 1$ соответственно. В дальнейшем для обозначения массы химических соединений будет использоваться обозначение $mass(\cdot)$. Вставка новой вершины определяется так, чтобы операции удаления и вставки были взаимно обратными. Допустима замена какого-либо ребра ((v, u) или (u, v)) на новую вершину w и два ребра ((v, w) и (w, u) или (u, w) и (w, v) соответственно), а также добавление висячей вершины w с любой положительной массой и с любым направлением нового ребра, но с соответствующим уменьшением $M(v)$ на $mass(H)$ (для ребра (v, w)) или $mass(OH)$ (для ребра (w, v)). Тем не менее, если $M(v) \leqslant mass(H)$ или $M(v) \leqslant mass(OH)$, то соответствующая вставка невозможна. Простую замену массы вершины я также буду называть *наивной модификацией*.

Определение 4. Если ППС-граф $G' = (V', E', M')$ получен из $G = (V, E, M)$ с помощью одной модификации, то *масса* этой *модификации* есть $\sum_{v' \in V'} M'(v') - \sum_{v \in V} M(v)$.

Определение 5. *k-кандидатом* называется пара $(G, \{(v_i, m_i)\}_{i=1}^{k})$, где $G = (V, E, M)$ — ППС-граф, $v_i \in V$, а m_i — вещественные числа.

Определение 6. ППС-граф G' будем называть *(наивным) k-вариантом* ППС-графа G, если G и G' не изоморфны, но G' может быть получен из G с помощью ровно k (наивных) модификаций, причём меньшего числа (наивных) модификаций недостаточно.

Другими словами, k-кандидаты — это ППС-графы вместе с предлагаемым списком наивных модификаций — парами (v_i, m_i), указывающими, массы каких вершины нужно изменить и на сколько. Именно 1-кандидаты являются результатом работы алгоритма VarQuest, а 2-кандидаты — результатом работы моего алгоритма. Алгоритм VarQuest предлагает делать именно наивные модификации, тем не менее по паре (v, m) можно осуществлять соответствующие вставку или удаление, если удаётся добиться данной массы модификации. По вершине v и массе m можно определить, какие модификации возможны в данном случае:

1. удаление v с $indeg(v) = outdeg(v) = 1$: $m = -M(v)$;

2. удаление v с $indeg(v) = 1$ и $outdeg(v) = 0$: $m = mass(H) - M(v)$;

3. удаление v с $indeg(v) = 0$ и $outdeg(v) = 1$: $m = mass(OH) - M(v)$;

4. вставка вершины вместо ребра: $m > 0$;

5. вставка висячей вершины: $m > -mass(H) > -M(v)$ или $m > -mass(OH) > -M(v)$;

6. замена: $m > -M(v)$.

Одновременно возможно выполнение максимум четырёх из перечисленных условий, поэтому по каждому k-кандидату может быть получено (с помощью применения соответствующих модификаций в указанном порядке) не более, чем 4^k различных k-вариантов.

В дальнейшем я буду работать только с (наивными) 1- и 2-вариантами. В случаях, когда речь будет идти об изоморфизме пептидных природных соединений, их модификациях или вариантах, то будут подразумеваться соответствующие свойства их графов.

1.3. Задачи идентификации пептидных природных соединений

Сформулируем строго задачи стандартной и вариативной идентификации ППС по базе данных. Даны множество известных пептидных природных соединений $\mathcal{P} = \{P_j\}_{j=1}^N$ (база данных ППС) и масс-спектр S, порождённый неизвестным ППС P. ППС $P_j \in \mathcal{P}$ является верным решением задачи стандартной идентификации спектра S по базе $\mathcal{P}$, если $G(P_j)$ изоморфен $G(P)$.

Аналогично k-кандидат $(G(P_j), \{(v_i, m_i)\}_{i=1}^k)$, $P_j \in \mathcal{P}$ является верным решением задачи вариативной идентификации спектра S по базе $\mathcal{P}$, если существует такая последовательность ППС-графов $G(P_j) = G_0, G_1, \ldots, G_k$, что G_i получается из G_{i-1} с помощью модификации массой m_i при вершине v_i, и G_k изоморфен $G(P)$.

1.4. Алгоритм вариативной идентификации пептидных природных соединений VarQuest

Алгоритм VarQuest [4] действует следующим образом. Для данного спектра S выбирается короткий список $CandidatePeptides(S) \subseteq \mathcal{P}$ из базы данных ППС. Затем с помощью под-алгоритма VariableScorer для каждого ППС $P \in CandidatePeptides(S)$ определяется предполагаемая позиция $v \in V$ модификации и её масса m (здесь V — множество вершин ППС-графа $G(P)$). Для соответствующего наивного 1-варианта G' вычисляется $s = Score(S, G')$, после чего в каждом случае находится p-значение — величина, отражающая статистическую значимость величины s для данных S и $G(P)$ [5]. Наконец, ППС P с наименьшим p-значением среди всех $P \in CandidatePeptides(S)$ возвращается алгоритмом как наиболее вероятное соединение, породившее спектр S.

Для дальнейшего не важно, как VarQuest определяет $CandidatePeptides(S)$ и как вычисляется p-значение, но понимание работы VariableScorer понадобится для реализации вариативной идентификации с двумя модификациями.

По тандемному масс-спектру S можно с высокой точностью определить массу его ori-ППС P. Поэтому, если предположить, что $G(P)$ является 1-вариантом $G(P_j)$ для некоторого $P_j \in \mathcal{P}$, то можно однозначно восстановить массу соответствующей модификации. Определив массу модификации m, VariableScorer перебирает все вершины ППС-графа $G(P_j) = (V, E, M)$ в поисках наиболее подходящей для модификации. Для каждой вершины $v \in V$ с помощью замены $M(v)$ на $M(v) + m$, если это допустимо, строится наивный 1-вариант G_v ППС-графа $G(P_j)$. Для каждого такого G_v вычисляется $Score(S, G_v)$ — величина, отражающая, насколько спектр S подходит ППС-графу G_v. На этом этапе алгоритм существенно опирается на масс-спектр. Наконец, среди всех $\{G_v \mid v \in V\}$ выбирается лучший 1-вариант $G' = \arg\max_{v \in V} Score(S, G_v)$. Таким образом, по масс-спектру S и пептидному природному соединению P_j VariableScorer находит наиболее подходящего 1-кандидата $(G(P_j), \{(v, m)\})$.

2. Сравнение ППС-графов

2.1. Проверка решений задач идентификации ППС

Прежде чем приступать к задаче вариативной идентификации, выясним, как мы можем проверять её решения. Для результатов реальной масс-спектрометрии, направленной на выяснение химического состава природных образцов, нужно знать ППС, породившее спектр, что уже невозможно без дополнительных экспериментов. Тем не менее, можно отобрать масс-спектры, которые были целенаправленно получены из конкретных пептидных природных соединений, и, зная правильный ответ, проверять на них результативность алгоритмов идентификации, что и было сделано.

Задача вариативной идентификации кажется трудной. На первый взгляд, даже проверка её решения алгоритмически нетривиальна. Рассмотрим более простую задачу стандартной идентификации. Проверка её решения по определению эквивалентна проверке изоморфизма ППС-графов. Ясно, что к этой задаче сводится задача проверки изоморфизма обычных ориентированных графов, для которой на данный момент неизвестно полиномиального от числа вершин в графе алгоритма. Для сведения достаточно присвоить всем вершинам равные массы.

Для всякого k-кандидата $(G(P_j), \{(v_i, m_i)\}_{i=1}^{k})$ существует не более 4^k различных с точностью до изоморфизма соответствующих ППС-графов последовательностей
$G(P_j) = G_0, G_1, \ldots, G_k$, где G_i получается из G_{i-1} с помощью модификации массой m_i при вершине v_i (в разделе 1.2 описано, как по массе определить, какие именно модификации можно делать). Таким образом, если считать k ограниченным (в нашем случае $k \leqslant 2$), то проверка решения задачи вариативной идентификации сводится к многократной проверке изоморфизма ППС-графов.

2.2. Структурные особенности ППС-графов

На основе вышесказанного можно было бы ожидать, что проверка результатов вариативной идентификации окажется вычислительно сложной, но на самом деле особенности строения ППС позволяют сделать это эффективнее. Нам нужно производить проверку изоморфизма двух ППС-графов, один из которых соответствует пептидному природному соединению из PNPdatabase. Анализируя ППС-графы в PNPdatabase, мне удалось выявить, что большинство из них имеют одну из следующих структур (направление рёбер и петли не учитываются, кроме того, здесь и далее подразумевается, что каждый упоминаемый путь в графе состоит хотя бы из одного ребра и, соответственно, хотя бы двух вершин):

1. *Линейная*: дерево, состоящее из одного простого пути, проходящего по всем вершинам;

2. *Циклическая*: граф, состоящий из одного простого цикла, проходящего по всем вершинам;

3. *Ветвь-циклическая*: граф содержит один простой цикл, из некоторых вершин которого выходят простые непересекающиеся между собой пути (не более одного из каждой вершины);

4. *Трёхлучевая звезда*: дерево, состоящее из трёх простых путей, выходящих из одной вершины, и больше непересекающихся;

5. *Гантель*: два непересекающихся цикла, соединённых простым путём.

Более того, каждый из ori-ППС отобранных мною спектров (см. раздел 3.1) имел одну из перечисленных выше структур. Количество ППС-графов в PNPdatabase каждого типа отображено в таблице 1.

2.3. Эффективный алгоритм проверки ППС-графов на изоморфность

Понятно, что изоморфные ППС-графы должны иметь одинаковую структуру, поэтому я предлагаю следующий алгоритм проверки изоморфизма ППС-графов. Сначала определяются структуры обоих ППС-графов, и, если они различны или в ППС-графах разное число вершин, то ППС-графы не изоморфны. Иначе, с учётом структуры перебираются возможные биекции между вершинами. Сначала определяются образы для некоторых ключевых вершин, а затем обходом графа в глубину из ключевых вершин — образы остальных вершин. Когда биекция построена, остаётся проверить равенство в пределах погрешности масс соответствующих вершин и направлений всех рёбер.

Обозначим через $degs_i(G)$ — число вершин в графе G степени i. Следующая теорема дает эффективный способ алгоритмического определения структуры ППС-графа.

Теорема 1. *Пусть $G = (V, E)$ — связный неориентированный граф без петель, $|V| \geqslant 2$. Его структура*

(a) линейная $\Leftrightarrow degs_1(G) = 2$ и $degs_2(G) = |V| - 2$;

(b) циклическая $\Leftrightarrow degs_2(G) = |V|$;

(c) ветвь-циклическая $\Leftrightarrow degs_1(G) = degs_3(G) > 0, \sum_{i=1}^{3} degs_i(G) = |V|$ и все вершины степени три лежат в одном простом цикле;

(d) трёхлучевая звезда $\Leftrightarrow degs_1(G) = 3, degs_2(G) = |V| - 4$ и $degs_3(G) = 1$;

(e) гантель $\Leftrightarrow degs_2(G) = |V| - 2, degs_3(G) = 2$ и существует ровно один простой путь между вершинами степени 3.

Доказательство. Все следствия слева направо вытекают из определения структур. Докажем утверждения справа налево.

(a) Запустим обход графа из вершины степени 1. Мы пройдём по некоторому пути через вершины степени 2 и остановимся в другой вершине степени 1. Поскольку в графе нет вершин степени 3, то пройденный путь есть целая компонента связности, но G связен, следовательно весь граф есть один простой путь.

(b) Аналогично предыдущему пункту, только остановим обход, когда вернёмся в стартовую вершину.

(c) Пусть $degs_1(G) = degs_3(G) = k > 0$, тогда $degs_2(G) = |V| - 2k$. Вычислим число рёбер в графе: $2|E| = 1 \cdot k + 2 \cdot (|V| - 2k) + 3 \cdot k = 2|V|$. То есть в графе G рёбер на единицу больше, чем в его остовном дереве. Следовательно, в G ровно один простой цикл, и по условию все вершины степени 3 лежат в нём. Далее, чтобы убедиться в ветвь-циклической структуре графа, достаточно снова запустить обход G из каждой висячей вершины, останавливая его при попадании в вершину степени 3. В силу связности, обход не может закончиться в другой висячей вершине.

(d) Аналогично предыдущим пунктам запускаем обход графа из каждой висячей вершины и останавливаем его при попадании в вершину степени 3.

(e) В очередной раз делаем обходы графа G, начиная и заканчивая в вершинах степени 3.

□

Определение структуры ППС-графов, построение одной биекции, проверка масс вершин и направлений рёбер — каждая из этих операций может быть выполнена за линейное время, но общая временная сложность также зависит от числа биекций, которые будут перебираться. Пусть в сравниваемых ППС-графах по n вершин. Ниже для каждого типа структуры перечислены ключевые вершины, с которых начинается построение биекций, и число проверяемых биекций:

1. Линейная: ключевые вершины — степени 1 (концы пути), число биекций — 2;
2. Циклическая: ключевая вершина — одна любая, число биекций — $2n$;
3. Ветвь-циклическая: ключевая вершина – одна любая из цикла, число биекций — $\leqslant 2n$;
4. Трёхлучевая звезда: ключевая вершина — степени 3, число биекций — $\leqslant 6$ $(= 3!)$;
5. Гантель: ключевые вершины — степени 3 (точки соединений циклов с путём), число биекций — $\leqslant 8$ $(= 2 \cdot 2 \cdot 2$, два способа совместить пути, и по два варианта совмещения циклов после этого).

Итак, получается итоговая оценка вычислительной сложности предложенного мною алгоритма проверки изоморфизма ППС-графов (см. таблицу 1). Отмечу, что для моих целей было особенно важно иметь эффективный способ проверки ППС-графов на изоморфность, поскольку к ней сводится вся проверка результатов вариативной идентификации. Для каждого масс-спектра как алгоритм VarQuest, так и мой алгоритм, могут находить сотни разных кандидатов, каждого из которых необходимо проверять.

Таблица 1: Количество ППС-графов в PNPdatabase в зависимости от структуры. Сложность проверки изоморфизма.

	# ППС в PNPdatabse	# ori-ППС	Временная сложность проверки изоморфности
Линейный	1395	78	$O(n)$
Циклический	1347	196	$O(n^2)$
Ветвь-циклический	1452	70	$O(n^2)$
Трёхлучевая звезда	165	3	$O(n)$
Гантель	66	6	$O(n)$
Остальные	659	0	
Всего	5084	353	

3. Методы

3.1. Данные

3.1.1 Спектры пептидных природных соединений из GNPS

Для проверки разработанных алгоритмов на корректность, точность и полноту, необходимы аннотированные масс-спектры пептидных природных соединений, т.е. спектры, про которые достоверно известно его ori-ППС. В качестве источника таких данных была выбрана база Global Natural Products Social Molecular Network (GNPS) [2]. Несмотря на огромное число тандемных масс-спектров, размещённых в GNPS и доступных исследователям по всему миру, лишь немногие из них имеют подтверждённое вещество, породившее спектр. Кроме того, в данной работе я сфокусирован только на пептидных природных соединениях, в то время, как в GNPS собираются данные, относящиеся ко всем классам природных соединений.

Из базы GNPS было извлечено более 11800 аннотированных спектров. Среди них были отобраны спектры пептидных природных соединений высокого разрешения, то есть с погрешностью определения масс фрагментов не более 0.02 Da. В итоге осталось 353 аннотированных масс-спектра.

3.1.2 База пептидных природных соединений

В качестве базы пептидных природных соединенй была взята PNPdatabase из [4]. Дополнительно к ней были добавлены все ori-ППС отобранных масс-спектров. Общий размер PNPdatabase составил 5 084 ППС.

3.2. Алгоритм идентификации спектра по ППС-графу с двумя модификациями

Ниже представлен VarScorer2 — придуманный мной алгоритм вариативной идентификации спектра с двумя возможными модификациями, масса одной из которых принадлежит конечному множеству, заданному пользователем.

Итак, для данных масс-спектра S, ППС-графа $G = (V, E, M)$ и списка из фиксированных масс модификаций Δ, перебираются всевозможные вершины графа $v \in V$ и $\delta_{fixed} \in \Delta$. Для каждой пары (v, δ_{fixed}) производится наивная модификация графа G (получается граф G'). После этого $VariableScorer$, являющийся частью оригинального алгоритма VarQuest, находит наиболее подходящее положение для второй модификации и вычисляет $s = Score(S, G')$.

Для фиксированных ППС-графа G и спектра S, p-значение пары (S, G') будет тем меньше, чем больше $Score(S, G')$, где G' – наивный вариант G. Поэтому внутри алгоритма 1 мы можем выбирать лучшего кандидата по s, оставив трудоёмкое вычисление p-значения на конец.

Алгоритм 1 VarScorer2

Вход: $(S, G = (V, E, M), \Delta)$

$s_{max} \leftarrow -\infty$
$res = None$
Для всех $v \in V$ **выполнять**
    Для всех $\delta_{fixed} \in \Delta$ **выполнять**
        $G' \leftarrow MakeReplacement(G, v, \delta_{fixed})$
        $u, \delta_{var}, s \leftarrow VariableScorer(S, G')$
        Если $s > s_{max}$ **тогда**
            $s_{max} \leftarrow s$
            $res \leftarrow ((v, \delta_{fixed}), (u, \delta_{var}))$
    Конец цикла
Конец цикла
Возвратить res

3.3. Определение допустимых модификаций

Ключевым шагом для улучшения качества вариативной идентификации стал созданный мною алгоритм ModAdmissibility для проверки кандидатов «на разумность». Его преимуществом является то, что ему не требуется никакая информация об ori-ППС и даже о масс-спектре.

Для того, чтобы понять идею ModAdmissibility, рассмотрим подробнее, как методы идентификации пептидных природных соединений применяются на практике. Смысл задачи идентификации в том, чтобы определить химическое строение некоторого ППС по его масс-спектру, при этом современные алгоритмы, такие как Dereplicator, VarQuest и мой VarScorer2, работают с ППС как с вершинно-взвешенными графами. Но даже если они находят верного k-кандидата для некоторого спектра, остаётся вопрос в том, как из соответствующего k-варианта (т.е. ППС-графа) восстановить химическое строение ППС. А именно по массе каждой вставленной или наивно модифицированной вершины необходимо восстановить аминокислоту, которая могла бы иметь соответствующую массу. Это и приводит к идее алгоритма ModAdmissibility: удостовериться в том, что по массам вершин k-варианта можно восстановить аминокислоты.

В качестве списка допустимых аминокислот я использовал все аминокислоты, которые встречаются в PNPdatabase. Здесь есть небольшая тонкость. Рассмотрим очень распространённую аминокислоту лейцин с формулой $C_6H_{13}NO_2$ и массой 131.09 Da. Лейцин имеет одну аминогруппу и одну карбоксильную группу, что позволяет ему образовывать две пептидные связи. Поэтому, в зависимости от положения лейцина в структуре ППС, соответствующий аминокислотный остаток может иметь следующий вид: $C_6H_{12}NO_2$ (масса 130.09 Da), $C_6H_{12}NO$ (масса 114.09 Da) или $C_6H_{11}NO$ (масса 113.08 Da). Мне не хотелось считать каждый из этих вариантов как отдельную аминокислоту, поэтому каждая аминокислота, встречающаяся в PNPdatabase, нормализовалась: к её формуле добавлялось необходимое число OH и H групп, в соответствии с количеством образованных ею пептидных связей. Таким образом, аминокислоты учитывались именно в нормализованной форме. Всего было обнаружено 1298 различных аминокислот.

Итоговый список был отсортирован по массам аминокислот, чтобы эффективнее искать аминокислоты с заданной массой. В таблице 2 представлены 20 наиболее часто встречаемых в PNPdatabase аминокислот.

Алгоритм 2 ModAdmissibility

Вход: **(**$G = (V, E, M), v, m$**)**

1: **Если** $\exists$ аминокислота с массой $M(v) + m + outdeg(v) * mass(H) + indeg(v) * mass(OH)$ **тогда**
2: **Возвратить** True

3: **Если** $\exists$ аминокислота с массой $m + mass(H_2O)$ **тогда**
4: **Возвратить** True

5: **Если** $M(v) + m = 0 \ \ \& \ \ indeg(v) = outdeg(v) = 1$ **тогда**
6: **Возвратить** True

7: **Если** $M(v) + m = mass(H) \ \ \& \ \ indeg(v) = 1 \ \ \& \ \ outdeg(v) = 0$ **тогда**
8: **Возвратить** True

9: **Если** $M(v) + m = mass(OH) \ \ \& \ \ indeg(v) = 0 \ \ \& \ \ outdeg(v) = 1$ **тогда**
10: **Возвратить** True

11: **Возвратить** False

Итак, алгоритм ModAdmissibility принимает на вход ППС-граф $G = (V, E, M)$, вершину $v \in V$, при которой предлагается делать модификацию, и массу модификации m. Затем он последовательно проверяет все допустимые модификации ППС-графов, и, если хотя бы одна из них возможна, масса модифкации признаётся допустимой, иначе — нет. Алгоритмом 2 проверяются замена (строки 1-2), вставка (строки 3-4) и удаление (строки 5-10).

Таблица 2: Наиболее часто встречающиеся аминокислоты в PNPdatabase.

Формула	Название аминокислоты (предположительно)	Масса (Da)	# вхождений в PNPdatabase	Максимальное число связей	
				Исходящих	Входящих
$C_6H_{13}NO_2$	Лейцин	131.09	4249	1	1
$C_3H_7NO_2$	Аланин	89.05	2984	2	1
$C_5H_{11}NO_2$	Валин	117.08	2664	1	1
$C_4H_9NO_2$	-	103.06	2466	1	1
$C_4H_9NO_3$	Треонин	119.06	2073	2	1
$C_2H_5NO_2$	Глицин	75.03	1953	2	1
$C_5H_9NO_2$	Пролин	115.06	1873	2	1
$C_3H_7NO_2$	Серин	105.04	1731	2	2
$C_9H_{11}NO_2$	Фенилаланин	165.08	1081	2	1
$C_5H_{10}N_2O_3$	Глутамин	146.07	1035	1	1
$C_6H_{14}N_2O_2$	Лизин	146.11	836	2	1
$C_4H_8N_2O_3$	Аспарагин	132.05	795	3	2
$C_5H_9NO_4$	Глутаминовая кислота	147.05	741	2	2
$C_7H_{15}NO_2$	-	145.11	714	1	1
$C_4H_7NO_4$	Аспарагиновая кислота	133.04	704	2	2
$C_9H_{11}NO_3$	Тирозин	181.07	692	2	1
$C_2H_4O_2$	Уксусная кислота	60.02	606	0	1
CH_2O_3	Угольная кислота	62.00	484	0	2
$C_{11}H_{12}N_2O_2$	Триптофан	204.09	471	1	1
$C_6H_{14}N_4O_2$	Аргинин	174.11	446	2	1

4. Результаты

4.1. Идентификация спектров по PNPdatabase

Я провёл идентификацию каждого из 353 аннотированных масс-спектров с каждым ППС из PNPdatabase с помощью реализованного мной алгоритма VarScorer2. В качестве фиксированных масс модификаций были выбраны ± 14Da (метилирование), ± 28Da (диметилирование) и ± 16Da (гидроксилирование). Каждый 2-кандидат подвергался проверке на допустимость алгоритмом ModAdmissibility. По результатам идентификации и проверок для каждого спектра был сформирован итоговый список из 2-кандидатов, отсортированный по возрастанию p-значений.

4.2. Качество идентификации в условиях отсутствия ori-ППС и его 1-вариантов

Чтобы оценить качество разработанных алгоритмов, я смоделировал ситуацию, когда для данного масс-спектра S в базе известных пептидных природных соединений нет ни ori-ППС, ни его 1-вариантов. В этом случае алгоритмы Dereplicator и VarQuest не способны корректно идентифицировать спектр. Чтобы сымитировать эти условия, достаточно исключить ori-ППС и его 1-варианты из списка 2-кандидатов спектра S, если они в нём были. Кроме того, аминокислоты из таких ППС не учитывались алгоритмом ModAdmissibility при проверке кандидатов.

Для определения точности моего метода в предложенных условиях была вычислена доля спектров, для которых в топ-1 (-3, -5, -10) есть корректные 2-кандидаты. Результаты представлены в таблице 3. Ввиду особенностей подсчёта p-значений алгоритмом $VariableScorer$, очень часто для разных кандидатов с одинаковой структурой ППС-графа и равными $score$ их p-значения также равны. Из-за этого нельзя однозначно отсортировать их в списке всех кандидатов. Поэтому я вычислял точность результатов для «лучшей» (когда правильные кандидаты оказываются выше) и «худшей» (когда правильные кандидаты оказываются ниже) из сортировок. Соответствующие числа указаны через дефис. Если же кандидатов с равными p-значениями упорядочивать случайно, то точность будет лежать в указанных диапазонах и, скорее всего, заметно ближе к нижней границе, поскольку неверных кандидатов практически всегда заметно больше, чем верных.

Таблица 3: Число идентифицированных спектров и их доля в зависимости от количества проверяемых кандидатов (Топ-k) и параметров метода.

Фильтрация с помощью ModAdmissibility	Массы фиксированных модификаций	Идентификация по кандидатам из			
		Топ-1	Топ-3	Топ-5	Топ-10
Да	± 14	**59-86 (16.7-24.4%)**	**88-113 (24.9-32.0%)**	**105-121 (29.7-34.3%)**	123-135 (34.8-38.2%)
Да	$\pm 14, \pm 16, \pm 28$	53-86 (15.0-24.4%)	77-113 (21.8-32.0%)	**100-126 (28.3-35.7%)**	**124-142 (35.1-40.2%)**
Нет	± 14	51-74 (14.4-21.0%)	74-101 (21.0-28.6%)	89-119 (25.2-33.7%)	114-131 (32.3-37.1%)
Нет	$\pm 14, \pm 16, \pm 28$	47-73 (13.3-20.7%)	69-97 (19.5-27.5%)	81-110 (22.9-31.2%)	106-131 (30.0-37.1%)

Из результатов видно, что при использовании меньшего числа фиксированных масс

точность идентификации по небольшому числу лучших кандидатов выше, чем при использовании большего набора. Это можно объяснить тем, что при увеличении списка фиксированных модификаций возрастает количество ложно-положительных идентификаций, ввиду большего пространства поиска. Однако, поскольку на практике для идентификации масс-спектра достаточно хотя бы одного верного кандидата в Топ-k, то с увеличением k большее разнообразие фиксированных масс модификаций получает преимущество, что начинает быть заметно уже при k, равном 5, и более чётко — при $k = 10$.

Стоит отметить, что для некоторых масс-спектров, с которыми я работал, в PNPdatabase не было ни одного такого 2-варианта их ori-ППС, что одна из отличающих их масс модификаций принадлежит заданному мной списку высокочастотных масс модификаций. Это означает, что данные спектры не могут быть корректно идентифицированы с помощью моего подхода, а в списках их 2-кандидатов не будет ни одного верного экземпляра. В таблице 4 представлено распределение спектров в зависимости от количества 2-вариантов их ori-ППС в PNPdatabase, а также доля идентифицированных из них в зависимости от отсечки.

Таблица 4: Распределение спектров в зависимости от количества 2-вариантов их ori-ППС в PNPdatabase, а также доля (в %) идентифицированных из них. Общий итог **не** учитывает спектры без 2-вариантов.

# 2-вариантов	Фиксированные массы модификаций: ± 14					Фиксированные массы модификаций: $\pm 14, \pm 16, \pm 28$				
	Топ-1	Топ-3	Топ-5	Топ-10	# спектров	Топ-1	Топ-3	Топ-5	Топ-10	# спектров
0	0.0	0.0	0.0	0.0	125	0.0	0.0	0.0	0.0	107
1	64.1-66.7	69.2	69.2	71.8	39	46.3-58.5	56.1-63.4	58.5-65.9	63.4-70.7	41
2	29.4-41.2	50.0-52.9	50.0-52.9	52.9	34	28.1-31.2	43.8-50.0	50.0-53.1	53.1	32
3	29.0-48.4	41.9-51.6	54.8-61.3	61.3-64.5	31	25.9-37.0	40.7-51.9	48.1-51.9	48.1-55.6	27
4	17.4-26.1	26.1-43.5	30.4-43.5	47.8	23	34.8	34.8-39.1	34.8-43.5	43.5-47.8	23
5	14.3-42.9	50.0-57.1	57.1	57.1	14	42.9-57.1	42.9-57.1	57.1-64.3	64.3-71.4	14
6-10	3.5-15.8	8.8-31.6	24.6-33.3	35.1-42.1	57	1.5-20.6	5.9-32.4	23.5-42.6	42.6-51.5	68
11-20	25.9-37.0	48.1-59.3	55.6-70.4	63.0-85.2	27	8.6-34.3	28.6-48.6	40.0-51.4	48.6-60.0	35
$\geqslant 21$	0.0	0.0	0.0-33.3	66.7-100.0	3	0.0	16.7	16.7-33.3	50.0-66.7	6
Всего ($\geqslant 1$)	25.9-37.7	38.6-49.6	46.1-53.1	53.9-59.2	228	21.5-35.0	31.3-45.9	40.7-51.2	50.4-57.7	246

4.3. Вклад ModAdmissibility в вариативную идентификацию

Для оценки эффективности разработанного алгоритма ModAdmissibility я дополнительно выяснил его влияние на вариативную идентификацию алгоритмом VarQuest. Я действовал аналогично случаю с VarScorer2: производилась идентификация каждого аннотированного масс-спектра с каждым ППС из PNPdatabase. По её результатам для каждого спектра был сформирован список из 1-кандидатов, отсортированный по возрастанию p-значений. ori-ППС соответствующих масс-спектров из этих списков исключались. Точность идентификации вычислялась аналогично, результаты отображены в таблице 5.

Из результатов видно, что фильтрация кандидатов с помощью придуманного мной алгоритма ModAdmissibility даёт безоговорочное преимущество в вариативной идентификации. Кроме существенного отсечения неверных кандидатов, значительно уменьшается разброс, вызванный равенством p-значений разных кандидатов. Таким образом, результаты идентификации меньше зависят от случайности.

Таблица 5: Число спектров (и их доля) идентифицированных алгоритмом VarQuest в зависимости от использования алгоритма ModAdmissibility и количества проверяемых кандидатов (Топ-k).

Фильтрация кандидатов с помощью ModAdmissibility	Идентификация по кандидатам из			
	Топ-1	Топ-3	Топ-5	Топ-10
Да	**140-182 (39.6-51.6%)**	**194-229 (55.0-64.9%)**	**218-244 (61.8-69.1%)**	**250-257 (70.8-72.8%)**
Нет	74-165 (21.0-46.7%)	101-223 (28.6-63.2%)	119-234 (33.7-66.3%)	131-252 (37.1-71.4%)

В таблице 6 указано распределение всех 1-кандидатов в зависимости от их корректности и результата ModAdmissibility. Учитывались только 1-кандидаты с p-значением меньше 10^{-10}. Алгоритм ModAdmissibility в первую очередь был направлен на фильтрацию неправильных кандидатов, которых, как видно из таблицы, более чем в 13 раз больше, чем правильных. Поэтому, его точность равна $\frac{8724}{8724+11} \approx 0.999$, а полнота — $\frac{8724}{8724+6813} \approx 0.561$. Дополнительно отметим низкий коэффициент ложно-положительных результатов: $\frac{11}{11+1156} \approx 0.009$. Этот показатель важен, так как нам не хотелось бы терять и так немногочисленных верных кандидатов.

Таблица 6: Таблица сопряжённости: распределение всех 1-кандидатов в зависимости от их корректности и результата ModAdmissibility.

	Допустимый	Не допустимый	Всего
Верный	1156	11	1167
Не верный	6813	8724	15537
Всего	7969	8735	16704

Заключение

В данной работе я разработал метод вариативной идентификации пептидных природных соединений по масс-спектрам с учётом двух возможных модификаций. При этом одна модификация предполагается фиксированной, а вторая – произвольная. Фиксированная модификация может быть как одной из высокочастотных, так и некоторой специальной, заданной пользователем в соответствии с его интересами. Метод был протестирован на 353 спектрах из базы GNPS [2] и PNPdatabase из 5 084 ППС [4].

Дополнительно был реализован подход к фильтрации кандидатов, который позволяет заметно улучшить качество вариативной идентификации вне зависимости от числа предлагаемых модификаций. Кроме того, в целях оценки качества разработанных алгоритмов, была исследована задача проверки графов на изоморфность, и для графов, соответствующих пептидным природным соединениям, придуман алгоритм с квадратичной вычислительной сложностью.

Реализованный метод позволяет идентифицировать масс-спектры ранее неизвестных пептидных соединений, которые не могли быть идентифицированы современными алгоритмами, что, в свою очередь, ускоряет поиск новых антибиотиков, иммунодепрессантов и других лекарств.

Список литературы

[1] Li, J.W. & Vederas, J.C. *Drug discovery and natural products: end of an era or an endless frontier?* Science **325**, 161–165 (2009).

[2] Wang, M. et al. *Sharing and community curation of mass spectrometry data with global natural products social molecular networking.* Nat. Biotechnol. **34**, 828–837 (2016).

[3] Mohimani, H., Gurevich, A., Mikheenko, A. et al. *Dereplication of peptidic natural products through database search of mass spectra.* Nat. Chem. Biol. **13,** 30–37 (2017).

[4] Gurevich, A., Mikheenko, A., Shlemov, A. et al. *Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra.* Nat. Microbiol. **3,** 319–327 (2018).

[5] Mohimani, H., Kim, S. & Pevzner, P. A. *A new approach to evaluating statistical significance of spectral identifications.* J. Proteome Res. **12,** 1560–1568 (2013).