ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
“САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”
(СПбГУ)

Образовательная программа бакалавриата “Математика”

Отчет о практике

на тему

Ранжирование кандидатов при вариативной идентификации пептидных природных соединений по масс-спектрам

Выполнил студент 3 курса бакалавриата
группа 19.Б01-мкн
Вяткин Никита Сергеевич

Научные руководители:
Ст. науч. сотр., к.ф.-м.н.,
Гуревич Алексей Александрович,
Ст. науч. сотр., к.ф.-м.н.,
Тагирджанов Азат Мухаммедович

Санкт-Петербург
2022

Содержание

1. Введение

Пептидные природные соединения (пептидные соединения, ППС) — органические вещества, состоящие из аминокислот, соединенных пептидными связями. ППС — ценный источник новых лекарств, в частности антибиотиков, противовирусных препаратов и иммунодепрессантов. Наиболее распространенный 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 определенным соотношением m/z отбираются и расщепляются на более мелкие фрагменты ионов. Затем эти фрагменты вводятся во второй масс-спектрометр, который, в свою очередь, разделяет их по соотношению m/z и производит подсчёт доли каждого из фрагментов. Так получается *масс-спектр* вещества.

Известная задача в биоинформатике — нахождение в базе данных веществ того ППС, который сгенерировал определенный масс-спектр. Однако во многих случаях вещество отсутствует в базе данных, тогда как его *модифицированный вариант* (*вариант*) – присутствует. Для нас *модификация* ППС — это замена, вставка или удаление одной из его аминокислот. Идентификация (по масс-спектру) неизвестного пептидного соединения из его известных вариантов называется *вариативной идентификацией* (в отличие от стандартной *идентификации*, когда ППС присутствует в базе данных). Далее мы будем говорить только о вариативной идентификации и считать, что в базе данных нет пептидных соединений, порождающих рассматриваемые масс-спектры.

Для вариативной идентификации создано несколько алгоритмов, но качество их работы ещё далеко от идеала. Одним из лучших таких алгоритмов является VarQuest [1]. Для данного масс-спектра VarQuest проводит поиск в базе данных и выдает список из возможных *кандидатов* — известных пептидных соединений, с указанием аминокислоты, модификация которой может привести к ППС, породившему исходный спектр. При этом результаты ранжируются по p-value — величине, отражающей статистическую значимость найденного кандидата, которая подсчитывается алгоритмом [3]. Несмотря на естественность и теоретическое обоснование этого способа ранжирования, при таком подходе не учитывается ряд природных признаков кандидатов, в первую очередь — масса модификации и ее положение в структуре ППС. Поскольку разнообразие пептидных соединений и их модификаций очень велико, то для отдельно взятого масс-спектра VarQuest находит в среднем несколько десятков кандидатов, большая часть из которых являются неверными. Цель моей работы — улучшить ранжирование результатов поисковой выдачи алгоритма VarQuest.

2. Постановка задачи

Я формулирую свою задачу в рамках обучения ранжированию [4] в этом разделе.

2.1. **Обучение ранжированию.** Обучающие данные D состоят из множества запросов (масс-спектров) Q, каждый из которых имеет список кандидатов x_q. Каждый кандидат $x_{q,i}$ для спектра q имеет метку релевантности $y_{q,i} \in \{0, 1\}$. При этом $y_{q,i} = 1$, если из кандидата $x_{q,i}$ действительно получается ППС, породивший спектр q. Иначе $y_{q,i} = 0$. Таким образом:

$$D = \{(x_q, y_q) \,|\, q \in Q\}$$

Поскольку ранжирование кандидатов для каждого спектра производится независимо, далее я буду опускать нижний индекс q и использовать x и y для краткости.

Обучение ранжированию заключается в поиске модели ранжирования Φ, которая может предсказать оценки релевантности s для всех кандидатов:

$$s = \Phi(x)$$

В моем подходе модель ранжирования будет предсказывать оценку релевантности для каждого кандидата отдельно, не обращая внимания на других кандидатов для этого же спектра, т. е.

$$s = (s_i) = (\phi(x_i)) = \Phi(x)$$

После чего ранжирование кандидатов x_i производится по убыванию s_i. Такие модели немного уступают более общим в качестве ранжирования, но они проще в реализации и обучении.

Несмотря на то, что y_i принимает лишь два значения, s_i может быть любым вещественным числом. Так, бóльшие значения s_i говорят о бóльшей уверенности модели в корректности кандидата.

2.2. **Функции качества ранжирования.** Существует множество функций качества, используемых в задачах обучения ранжированию. Общим свойством этих функций является то, что они зависят от ранга и уделяют больше внимания релевантности кандидатов с наивысшим рейтингом.

Пусть m – общее число кандидатов для спектра q, а π – перестановка чисел от 1 до m, соответствующая какому-то ранжированию x. Т. е. кандидату $x_{\pi(1)}$ отдается высший приоритет, а $x_{\pi(m)}$ – низший.

Для оценки качества ранжирования я использовал две функции: MAP и $NDCG$ [4].

2.2.1. *Mean Average Precision (MAP).* Чтобы определить MAP, мы сначала должны определить *точность в позиции k (P@k)*:

$$P@k(q, \pi) = \frac{\sum_{i=1}^{k} y_{\pi(i)}}{k}$$

Проще говоря, $(P@k)(q, \pi)$ – доля релевантных кандидатов среди первых k. Затем, *Average Precision (AP)* определяется как

$$AP(q, \pi) = \frac{\sum_{k=1}^{m} P@k(q, \pi) \cdot y_{\pi(k)}}{\sum_{k=1}^{m} y_{\pi(k)}}$$

что является средней точностью в позиции по всем позициям релевантных кандидатов. Среднее значение AP по всем тестовым спектрам называется MAP.

Определим дополнительно *полноту в позиции k (R@k)*:

$$R@k(q, \pi) = \frac{\sum_{i=1}^{k} y_{\pi(i)}}{\sum_{i=1}^{m} y_{\pi(i)}}$$

Полнота показывает, какая часть релевантных кандидатов оказалась на первых k позициях. Кривая точности-полноты получается путем построения графика зависимости точности от полноты для всех k. Полезной особенностью AP является то, что она может быть визуализирована как площадь под кривой точности-полноты.

2.2.2. *Normalized Discounted Cumulative Gain (NDCG).* В то время как MAP в основном предназначена для задач с бинарной релевантностью, $NDCG$ может использоваться при ранжировании объектов с несколькими уровнями релевантности и имеет в своем определении явный коэффициент дисконтирования позиции. Более формально, *Discounted Cumulative Gain в позиции k* ($DCG@k$) определяется следующим образом:

$$DCG@k(q, \pi) = \sum_{i=1}^{k} G(y_{\pi(i)})\eta(i)$$

где $G(\cdot)$ – рейтинг кандидата (обычно используется $G(y_{\pi(i)}) = 2^{y_{\pi(i)}} - 1$), а $\eta(i)$ – коэффициент дисконтирования позиции (обычно используется $\eta(i) = 1/\log_2(i+1)$).

Поделив $DCG@k$ на его максимальное значение (которое достигается, если всех кандидатов упорядочить по y и обозначается как Z_k), мы получим другую функцию, называемую *Normalized Discounted Cumulative Gain* ($NDCG$). То есть:

$$NDCG@k(q, \pi) = \frac{1}{Z_k(q)} \sum_{i=1}^{k} G(y_{\pi(i)})\eta(i)$$

Понятно, что $NDCG$ принимает значения от 0 до 1. В случае, когда учитываются все кандидаты, вместо $DCG@m$ и $NDCG@m$ пишут просто DCG и $NDCG$ соответственно.

3. Методы

На этапе фрагментации ионов при тандемной масс-спектрометрии разрушаются в основном пептидные связи — химические связи, соединяющие аминокислоты, поэтому удобно считать аминокислоты неделимыми структурными единицами пептидных соединений. При таком подходе каждое ППС представляется в виде графа, вершины которого — аминокислоты, а ребра — соединяющие их пептидные связи.

3.1. **Представление и обработка данных.** Как отмечалось ранее, для каждого масс-спектра $q \in Q$ дан список кандидатов — пептидных соединений с модификациями (см табл. 1). Про ППС полностью известны его атомное строение и, как следствие, атомная масса (mass) и количество аминокислот в составе (num_aa). Кроме положения модификации (mod_node, формула аминокислоты и ее масса), известна ее масса (mod_mass_shift). Также для каждого кандидата алгоритм VarQuest возвращает его предполагаемую релевантность в виде двух параметров: score и p-value. Для удобства p-value было переведено в логарифмическую шкалу (log_pvalue). Дополнительно для каждого кандидата я подсчитал степень вершины (deg), которую предлагается модифицировать. Наконец, y_true — релевантность кандидата.

Масс-спектры, у которых все кандидаты имеют одинаковую релевантность (все правильные или все неправильные), были исключены из данных. Значение функций качества на них никак не зависило бы от ранжирования. После этого данные были разделены на тренировочную и тестовую выборки в соотношении 80 : 20.

3.2. **Вычисление частотности масс модификаций.** Массу модификации не эффективно использовать как признак для моделей машинного обучения в чистом виде. Каждая возможная в ППС модификация соответствует некоторому изменению в атомном составе, поэтому их разнообразие определяется разнообразием последних.

scan	mass	num_aa	score	log_pvalue	mod_node	mod_mass_shift	deg	mod_mass_freq	y_true
823	717.410	6	10.0	-44.530857	(C9H9NO, 147.068)	35.0086	2	-6.003059	0
823	793.463	7	10.0	-41.185407	(C2H3NO, 57.0215)	-41.0439	2	0.774721	0
823	638.351	5	6.0	-31.967684	(C6H10O2, 114.068)	114.0670	2	1.733681	1
823	831.501	7	8.0	-31.360110	(C6H11NO, 113.084)	-79.0820	2	-8.008483	0
823	524.283	4	5.0	-27.344828	(C6H13O6, 181.071)	228.1350	1	-3.109666	1
823	701.425	6	6.0	-24.071465	(C6H11NO, 113.084)	50.9936	2	-1.494799	0
823	676.368	6	6.0	-24.071465	(C5H8O2, 100.052)	76.0504	2	2.109222	0
823	618.317	6	6.0	-24.071465	(C3H5NO, 71.0371)	134.1020	2	-8.008483	0
823	819.453	7	6.0	-23.687352	(C3H5NO, 71.0371)	-67.0344	2	-0.297757	0
823	829.495	8	5.0	-23.120562	(C6H11NO, 113.084)	-77.0762	2	-7.995239	0

Таблица 1. Пример списка кандидатов для масс-спектра под номером 823

Напомню, что пептидное соединение p_1 называется *(модифицированным) вариантом* пептидного соединения p_2, если p_2 получается из p_1 при помощи одной модификации. *Модифицированная пара* — ППС и его вариант.

Если ППС p_1 является модифицированным вариантом ППС p_2 и отличается от него по массе на $+\delta$, то и p_2 – модифицированный вариант p_1 с массой модификации $-\delta$. Таким образом, если мы хотим понять насколько распространена определенная масса модификации, разумно обращать внимание только на ее абсолютную величину. Далее я буду считать, что разница в массе в каждой модифицированной паре положительна.

Обработав доступную мне базу данных ППС (которая была получена путем объединения всех ППС из AntiMarin [5], DNP [6], MIBiG [7] и StreptomeDB [8], и состояла из 5021 пептидного соединения) я нашел все модифицированные пары и абсолютную величину разницы масс в них.

Через $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ я буду обозначать плотность нормального распределения со средним μ и дисперсией σ^2, а через $\mathcal{U}(\cdot \mid [a,b])$ — плотность равномерного распределения на отрезке $[a,b]$. Пусть m — масса модификации, частоту наблюдения которой необходимо оценить, V — случайная величина, равная изменению атомного состава для данной модификации, v — все возможные значения V, наблюдаемые в базе данных, а $mass(v)$ – масса модификации, соответствующая v. Наконец, я использовал сглаживание к равномерному распределению (сглаживание Лапласа), чтобы избежать масс модификаций с нулевой частотой. С учетом всего вышесказанного, частота массы модификации $p(m)$ оценивается по формуле:

$$p(m) = \alpha \mathcal{U}(m \mid [0, MaxMod]) + (1-\alpha) \sum_v p(m \mid V = v) p(V = v)$$

$$= \alpha \frac{1}{MaxMod} + (1-\alpha) \sum_v \mathcal{N}\left(m \mid mass(v), \delta^2\right) \frac{n(v)}{N}$$

где $\alpha = 5 \cdot 10^{-4}$ — коэффициент сглаживания (выбор α – см. далее), $MaxMod$ — максимально возможная абсолютная величина массы модификации (алгоритм VarQuest находит только кандидатов с разницей в массе не более 300Da, поэтому я использовал $MaxMod = 300$), δ — погрешность определения массы модификации в соответствии с той точностью, с которой современные масс-спектрометры могут определять массу молекул (я использовал $\delta = 0.01$Da), $n(v)$ — кол-во найденных модифицированных пар с разницей

в атомном составе равной v, а N — общее число найденных пар. Частоты получались близкими к нулю, поэтому я перевел их в логарифмическую шкалу (mod_mass_freq).

3.3. **Отбор и обработка признаков.** Для отбора признаков были посчитаны попарные корреляции между признаками (см. рис. 1):

Сильнее всего с метками релевантности (y_true) скоррелированы логарифм p-value, score и частотность массы модификации. Но от использования score я решил отказаться, потому что он сильно коррелирует с p-value, а совместное использование таких признаков может привести к нестабильности обучения. Оба выбранных признака были нормализованы.

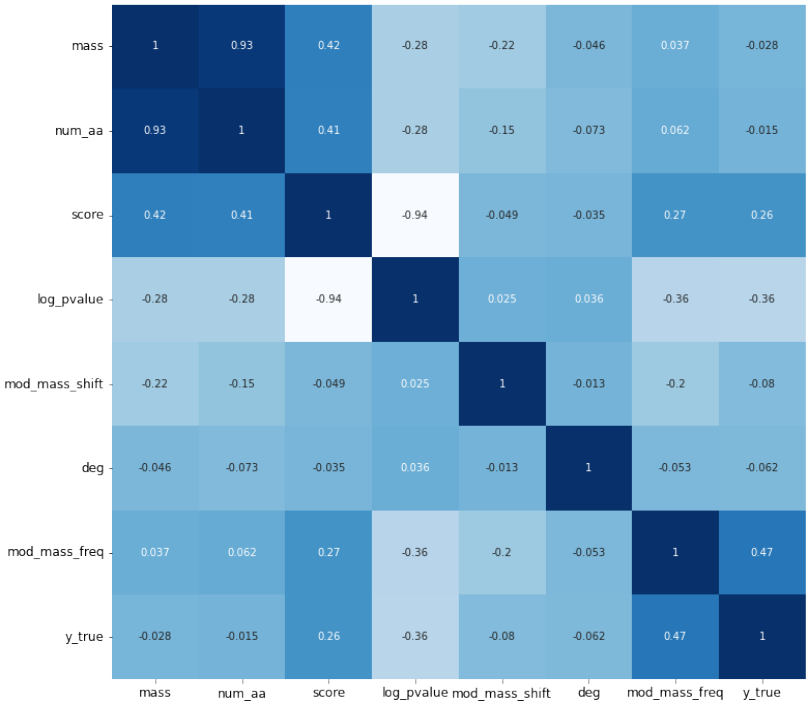

Рис. 1. Корреляции признаков

3.4. **Обучение модели.** В качестве модели машинного обучения для предсказания оценки релевантности я выбрал логистическую регрессию. Во-первых, это хорошая модель для задач бинарной классификации (напомню, что метки релевантности у кандидатов принимают значения лишь 0 или 1), во-вторых, логистическая регрессия, в отличие от таких моделей как SVM и дерево принятия решений, позволяет получать оценки вероятностей принадлежности к целевому классу, а значит лучше подходит для сортировки кандидатов.

3.4.1. *Поиск гиперпараметров.* В моей модели было три гиперпараметра: коэффициент сглаживания частот масс модификаций (α), способ регуляризации (L1 или L2) и коэффициент регуляризации. Их я подбирал с помощью k-fold кросс-валидации ($k = 5$).

Лучшие результаты при валидации были достигнуты на значениях: $\alpha = 5 \cdot 10^{-4}$, способ регуляризации – L2 с коэффициентом $5 \cdot 10^3$. При этом параметр α перебирался в диапазоне $[1.0 \cdot 10^{-5}, 1.0]$, а коэффициент регуляризации – $[0.1, 1.0 \cdot 10^4]$.

4. Результаты

4.1. **Сравнение значений функций качества.** На тестовых данных я сравнил ранжирование своей модели со стандартным ранжированием алгоритма VarQuest. В таблице 2 отражены значения функций качества: MAP, $NDCG$ и $NDCG@k$ для $k = 1, 3, 5$. Заметим, что $NDCG@1$ — суть доля спектров, для которых первый кандидат правильный.

	MAP	NDCG	NDCG@1	NDCG@3	NDCG@5
Стандартное ранжирование (по p-value)	71.38%	83.57%	75.24%	74.11%	71.92%
Ранжирование, полученное с помощью логистической регрессии	79.57%	89.72%	86.72%	84.80%	83.61%

Таблица 2. Значения функций качества

На рис. 2 изображены усредненные по всем тестовым данным кривые точности-полноты. Глядя на них, можно заметить существенное снижение точности при полноте больше 0.5. Вероятно, это особенность наших данных. В среднем для каждого спектра из всех релевантных кандидатов есть половина простых и легко идентифицируемых, которые оказываются на верхних позициях. Они имеют распространенные массы модификаций и достаточно маленькое значение p-value. Вторая половина правильных кандидатов, наоборот, с не самыми популярными массами модификаций и невыдающимися значениями p-value. Из-за этого их трудно отличить от нерелевантных кандидатов.

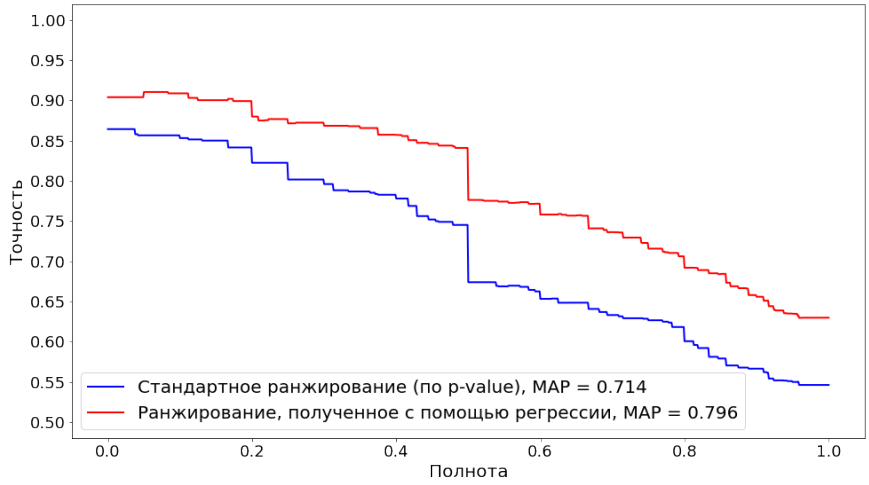

Рис. 2. Усредненные кривые точности-полноты (точность по оси Y в пределах от 0.5 до 1.0)

4.2. **Время работы.** Внедрение обученной модели ранжирования в качестве финального шага алгоритма VarQuest практически не повлияет на время работы последнего. Предсказание меток релевантности с помощью логистической функции требует константного времени для каждого кандидата. После чего необходимо пересортировать результаты за $O(n \log n)$ времени или выбрать фиксированное число лучших за $O(n)$ времени (здесь n — количество кандидатов для данного масс-спектра).

4.3. **Частотности модификаций.** Неудивительной была большая разница в частотности масс модификаций (см. таблицу 3; положительный/отрицательный индекс у химического элемента равен тому, насколько увеличилось/уменьшилось кол-во соответствующих атомов в составе в результате модификации). Например, более чем четверть всех найденных модифицированных пар отличается на 14Da, что соответствует метилированию — добавлению группы CH_2, в то время как модификаций массой 93Da или 116Da не было обнаружено вовсе.

Разница в атомном составе	Масса модификации (Da)	Сколько раз встретилась (доля)
CH_2	14.016	2014 (25.31%)
C_2H_4	28.031	624 (7.84%)
O	15.995	563 (7.07%)
H_2	2.016	240 (3.02%)
C_3H_6	42.047	219 (2.75%)
CH_2O	30.011	173 (2.17%)
$N_{-1}OH_{-1}$	0.984	172 (2.16%)
C_3H_{-2}	33.984	151 (1.90%)
C_2H_2	26.016	112 (1.41%)
C_4H_8	56.063	92 (1.16%)

Таблица 3. Самые популярные модификации

5. Заключение

В этой статье я презентовал подход, улучшающий ранжирование кандидатов при вариативной идентификации пептидных природных соединений по масс-спектрам с помощью алгоритма VarQuest. По сравнению со стандартным ранжированием мне удалось добиться улучшения на 6–11%, в зависимости от способа измерения качества. Реализованный метод позволяет быстрее и точнее идентифицировать масс-спектры ранее неизвестных пептидных соединений, что, в свою очередь, ускоряет поиск новых антибиотиков, иммунодепрессантов и других лекарств.

Список литературы

[1] 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).

[2] 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).

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

[4] Tie-Yan Liu. *Learning to Rank for Information Retrieval.* Springer Berlin, Heidelberg. Vol. 3, No. 3, 225–331 (2009).

[5] Blunt, J., Munro, M. & Laatsch, H. *AntiMarin Database* (Univ. Canterbury, Christchurch, and Univ. Gottingen, Gottingen, 2007); https://www.scienceopen.com/document?vid=03a1a98e-434c-4255-a287-5a900f59d024

[6] Gozalbes, R. & Pineda-Lucena, A. *Small molecule databases and chemical descriptors useful in chemoinformatics: an overview.* Comb. Chem. High T. Scr. **14,** 548–458 (2011).

[7] Medema, M. H. et al. *Minimum information about a biosynthetic gene cluster.* Nat. Chem. Biol. **11,** 625–631 (2015).

[8] Lucas, X. et al. *StreptomeDB: a resource for natural compounds isolated from Streptomyces species.* Nucleic Acids Res. **41,** D1130–D1136 (2013).