Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

на правах рукописи

Сулимов Павел Андреевич

ОБУЧЕНИЕ ГЕНЕРАТИВНЫХ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ ДЛЯ РАСПОЗНАВАНИЯ ДАННЫХ МАСС-СПЕКТРОМЕТРИИ

РЕЗЮМЕ

диссертации на соискание ученой степени
кандидата компьютерных наук

Москва — 2020

Диссертационная работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель:	Аттила Кертес-Фаркаш, PhD, доцент, факультет компьютерных наук Национального исследовательского университета «Высшая школа экономики»

1 Введение

Масс-спектрометрия используется для изучения и идентификации молекул в биологических образцах. Масс-спектрометр генерирует своего рода "отпечатки пальцев"молекул, называемые масс-спектрами, которые затем проходят процедуру идентификации или аннотирования исходных молекул, которые могли бы генерировать данные масс-спектры.

Масс-спектрометрия привлекла внимание в различных областях, включая молекулярную биологию, криминалистику, фармацевтическую промышленность, медицину и т.д. Например, при анализе содержимого объектов окружающей среды масс-спектрометрия может использоваться для проверки продуктов питания и напитков на предмет использования вредных химикатов. Также с помощью масс-спектрометров можно проводить анализ почвы, чтобы оценить количество пестицидов или гормонов, используемых при ее обработке. В криминалистике масс-спектрометрия может быть использована для подтверждения злоупотребления наркотиками, или же выявления взрывоопасных остатков и жидкостей для розжига при расследовании дел о поджогах. В фармацевтике определение структур лекарственных средств и метаболитов, а также поиск наличия метаболитов в биологических системах являются основными областями применения масс-спектрометрического анализа. В клинических исследованиях и разработке клинических препаратов масс-спектрометр используется для скрининга заболеваний, мониторинга медикаментозной терапии с целью изучения изменения белкового состава исследуемых клеток, а также выявления инфекционных агентов для таргетной терапии.

Эта диссертация посвящена идентификации белков в биологических образцах по данным масс-спектров, полученных с помощью тандемной масс-спектрометрии.

1.1 Акутальность исследования

С одной стороны, большое количество данных из масс-спектрометров было накоплено из-за резкого снижения стоимости устройств хранения данных в последние несколько лет. С другой стороны, недавняя разработка вычислительных устройств, таких как графические процессоры (GPU), позволяет исследователям в короткие сроки разрабатывать вычислительно эффективные и ресурсоемкие методы. Как следствие, в последнее время было опубликовано несколько статей с результатами применения методов глубокого обучения для аннотации данных масс-спектров, таких как MS2PIP [9], pDeep [64], Prosit [21], DeepNovo [48] и т.д.

1.2 Цели и задачи исследования

Целью данного исследования была разработка более точных методов идентификации масс-спектров. Мною был разработан новый метод, названный BoltzMatch, который основан на стохастической нейронной сети и может производить более точную аннотацию данных масс-спектров; более того, в отличие от других алгоритмов, представляющих собой «черный ящик», BoltzMatch интерпретируем. На момент написания этой диссертации, BoltzMatch до-

стиг современного уровня аннотации масс-спектров. Разработка метода BoltzMatch включала следующие задачи:

1. BoltzMatch требовал нового метода калибровки рейтингов (score), названного Tailor, для обеспечения того, чтобы итоговые рейтинги, полученные с его помощью, были нормализованы и, следовательно, сопоставимы между масс-спектрами.

2. Обучение BoltzMatch потребовало разработки нового метода регуляризации. Этот метод был назван диверсифицирующей регуляризацией. В исследовании также показано, что регуляризация помогает обучать произвольные глубокие стохастические нейронные сети.

3. В работе показано, что методы машинного обучения могут переобучаться, в результате чего происходит смещение оценок из-за переусложнения модели. Мною показано, что метод BoltzMatch не дает смещенных оценок в результате обучения.

1.3 Практическая значимость

Неправильная аннотация масс-спектра может привести к тому, что ученые-биологи и практики сделают ложные выводы о результатах экспериментов и, как следствие, принимут неверные решения, например, при выборе правильной лекарственной терапии. Поэтому важно разработать надежные и точные методы для аннотирования и идентификации масс-спектров, в независимости от протокола эксперимента.

1.4 Публикации

В результате моей исследовательской работы в аспирантуре было опубликовано 4 основных статьи, 3 из которых были опубликованы в журналах Q1. Рейтинг основан на Scopus и Web of Science.

Публикации повышенного уровня.

1. Sulimov P., Voronkova A., Kertész-Farkas A. Annotation of tandem mass spectrometry data using stochastic neural networks in shotgun proteomics. **Bioinformatics, Q1 журнал**, 2020, doi: https://doi.org/10.1093/bioinformatics/btaa206, 2020.

2. Sulimov P., Kertész-Farkas A. Tailor: A Nonparametric and Rapid Score Calibration Method for Database Search-Based Peptide Identification in Shotgun Proteomics. **Journal of Proteome Research, Q1 журнал**, 2020, doi: https://doi.org/10.1021/acs.jproteome.9b00736.

3. Danilova Y., Voronkova A., Sulimov P., Kertész-Farkas A. Bias in False Discovery Rate Estimation in Mass-Spectrometry-Based Peptide Identification. **Journal of Proteome Research, Q1 журнал**, 2019, doi: https://doi.org/10.1021/acs.jproteome.8b00991

Публикации стандартного уровня.

4. Sulimov P., Sukmanova E., Chereshnev R., Kertész-Farkas A. Guided Layer-Wise Learning for Deep Models Using Side Information. In: van der Aalst W. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2019. **Communications in Computer and Information Science, Q3 серия книг**, vol 1086. Springer, 2020, doi: https://doi.org/10.1007/978-3-030-39575-9_6

Доклады на конференциях и семинарах.

5. Guided Layer-wise Learning for Deep Models using Side Information, 8 международная конференция - Анализ Изображений, Сетей и Текстов, 17-19 июля 2019 года, Казань, Россия.

6. Generative probabilistic modelling of peptide-spectrum matching in tandem mass spectrometry, X международный форуме "Biotechnology: State Of The Art and Perspectives 25-27 февраля 2019 года, Москва, Россия.

7. Модификация ограниченной машины Больцмана для распознавания пептидов, Ежегодная межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е. В. Арменского, МИЭМ НИУ ВШЭ, 19 февраля 2018 года, Москва, Россия.

8. High-dimensional Generative Probabilistic Models for Peptide-spectrum Matching in Tandem Mass Spectrometry, II Российско-французский научный семинар "Большие данные и решения на их основе 12-13 октября 2017 года, Москва, Россия.

1.5 Содержание работы

Диссертация организована следующим образом. Глава 2 содержит краткий обзор метода масс-спектрометрии, научную дискуссию на предмет подготовки биологических проб, генерации данных, процесса идентификации масс-спектра с помощью поиска по базам данных и валидации результатов. В Главе 3 дается более подробное описание свойств функций скоринга и объясняется, почему их невыполнение может привести к уменьшению статистической мощности самой функции скоринга. Наконец, Глава 4 суммирует результаты, опубликованные в статьях, написанных в течение всего исследовательского процесса.

2 Масс-спектрометрия

2.1 Подготовка биологической пробы

В этой диссертации под биологическим образцом имеется ввиду небольшая аликвота, содержащая смесь нескольких белков, и главной целью является идентификация белковых последовательностей в образце [1, 47, 37, 49]. Образец может быть взят, например, из крови, или из раковой клетки при помощи биопсии; также в качестве биологического образца могут выступать небольшие молекулы, используемые бактериями для связи друг с другом и т.д.

Белки в образце сначала разрезаются с использованием ферментов, способствующих расщеплению, в результате чего образуются более мелкие молекулы, называемые пептидами. Проведение данной процедуры необходимо по следующим причинам:

1. уменьшить размер молекул, потому что современные приборы все еще не могу анализировать довольно крупные молекулы;

2. уменьшить сложность анализа.

Ферменты расщепляют белки в определенных местах, сама процедура называется разрезанием или расщеплением. Место расщепления зависит от фермента, и разные ферменты могут расщеплять в разных местах белка.

Наиболее популярным ферментом является трипсин, который расщепляет белковую последовательность после аминокислот аргинина (R) и лизина (K), если за ними не следует пролин (P) с C-конца [53]. Ферменты могут не расщеплять в определенных местах, если они недоступны из-за структуры белка.

2.2 Масс-спектрометр и генерация данных

Масс-спектрометр - это прибор, который принимает на вход аликвоту расщепленного образца, которая может быть в твердой, жидкой или газообразной фазе. Прибор производит набор масс-спектров, которые можно рассматривать как "отпечатки пальцев"молекул, подлежащие идентифицированию с помощью компьютерных программ.

Работа стандартного масс-спектрометра состоит из следующих этапов:

1. *Испарение*. Образец преобразуется в газообразную фазу.

2. *Ионизация*. Молекулы заряжаются положительно, заряженная молекула (катион) называется родительским (молекулярным) ионом. В основном, молекулы заряжаются либо путем удаления электрона(ов) с незначительным изменением соотношения массы/заряда, либо путем добавления одного или нескольких протонов, которые значительно изменяют соотношение массы/заряда пропорционально массе протона.

3. *Ускорение*. Ионизированные образцы ускоряются с помощью магнитного поля.

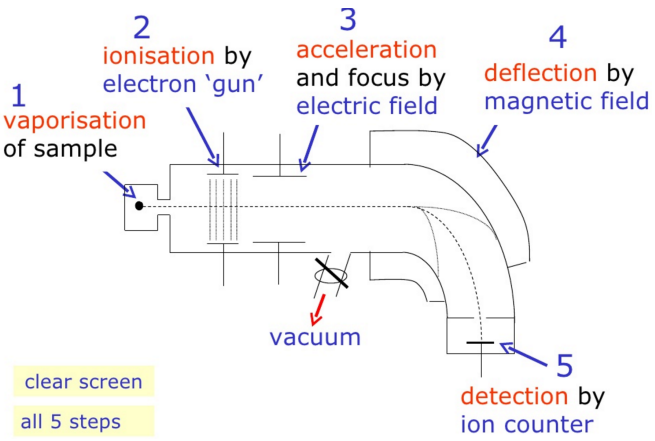

Рис. 1: Схема работы масс-спектрометра [7].

4. *Отклонение.* Пути распространения заряженных частиц отклоняются магнитом. Более тяжелые молекулы отклоняются меньше, более легкие - больше. На этой стадии заряженные частицы разделяются (сортируются) в соответствии с их отношением массы к заряду.

5. *Детекция.* Заряженные частицы регистрируются - в настоящее время это делается с достаточно высоким разрешением, чтобы различать изотопы; происходит подсчет ионов.

На Рисунке 1 представлена типичная схема устройства масс-спктрометра.

Одним из ключевых отличий между приборами является тип ионизации. Например, во время электронной ионизации [33, 28] электроны ударяются о молекулы и выбивают один электрон из облака электронов, делая молекулы положительно заряженными. При электроспрейной ионизации (ESI) молекулы оборачиваются в микрокапельки или покрываются водой или липидами, таким образом защищая молекулы пробы. Захваченные молекулярные ионы проходят через камеры высокого давления, где под давлением образец нагревается, и жидкая часть испаряется. В резуьтате чего только молекулярные ионы остаются и попадают в детектор. Ионизацию электроспреем можно сочетать с высокоэффективной жидкостной хроматографией (HPLC), и при таком подходе масс-спектрометр будет определять массы молекул, разделенных в аналитической колонке. Этот метод получил сокращение LC-MS (жидкостная хроматография с масс-спектрометрией) [51]. Идентификация белков в сложном растворе с использованием комбинации HPLC и масс-спектрометрии называется скорострельной протеомикой [2, 62, 19], которая в настоящее время широко используется.

Подход *тандемной масс-спектрометрии* (MS/MS) показан на Рисунке 2. Тандемные масс-спектрометры используются с «мягкими» методами ионизации, в которых не используются электроны или химические молекулы. При таком подходе первый масс-спектрометр анализирует все молекулярные ионы, а затем молекулярные ионы дополнительно фрагментируются в результате столкновений с молекулами инертного газа или лазерного излучения во втором масс-спектрометре. В итоге, каждый пептид характеризуется набором фрагментарных ионов, наряду с родительским ионом, и соответствующих им масс, называемым масс-

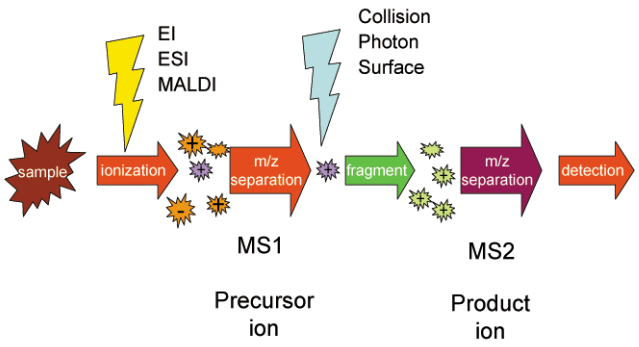

Рис. 2: Схематичное представление процесса тандемной масс-спектрометрии[5].

спектром.

На Рисунке 3 представлена теоретическая фрагментация данного пептида *leuenkephalin*. В пептиде амидная связь (CO-NH) является наиболее уязвимой, и потому это является наиболее вероятным местом, в котором пептид разделяется на две части, называемые $b-$ион и $y-$ион. Пептидный фрагмент, содержащий N-конец, называется $b-$ионом, тогда как другой фрагмент, содержащий C-конец, называется $y-$ионом. Сумма масс $b-$ и $y-$ионов равна массе исходного пептидного иона. Расстояние между двумя соседними ионами ($b-$ или $y-$) равно массе аминокислоты, которая разделяет данные два иона. Следовательно, определение аминокислот, соответствующих расстояниям в ряду фрагментарных $b-$ионов, позволит восстановить исходную пептидную последовательность (кроме последней аминокислоты), а аналогичное определение аминокислот, соответствующих расстояниям в ряду фрагментарных $y-$ионов, позволит восстановить исходную пептидную последовательность в обратном порядке (кроме первой аминокислоты).

На Рисунке 4 показана иллюстрация спектра, наиболее вероятно генерируемого пептидом **HEEIDLVSLEEFYK**. Вдоль оси абсцисс отмеряется отношение массы к заряду m/z (Th). Масса измеряется в Дальтонах (Da). Точность масс-спектрометра низкого разрешения находится в пределах 1 Da, точность масс-спектрометра высокого разрешения - до 0.02 Da, что означает, что он может регистрировать значения с точностью до $1/50$ массы протона. Значения оси Y соответствуют интенсивности, то есть отражают количество наблюдаемых фрагментарных ионов.

2.3 Идентификация масс-спектра на основе поиска по базе данных

В подходе поиска по базе данных наблюдаемые спектры $s_i \in S$, где S обозначает набор спектров, полученных прибором из одного эксперимента, аннотируются теоретическим пептидом с наибольшим рейтингом, найденным путем итеративного сопоставления s_i с базой данных собранных пептидов, составляя так называемую пару *теоретический пептид - настоящий спектр* (PSM), формально:

$$s_i \leftarrow \hat{h} = \arg\max_{h_j \in CP(s_i)} \phi(s_i, h_j). \tag{1}$$

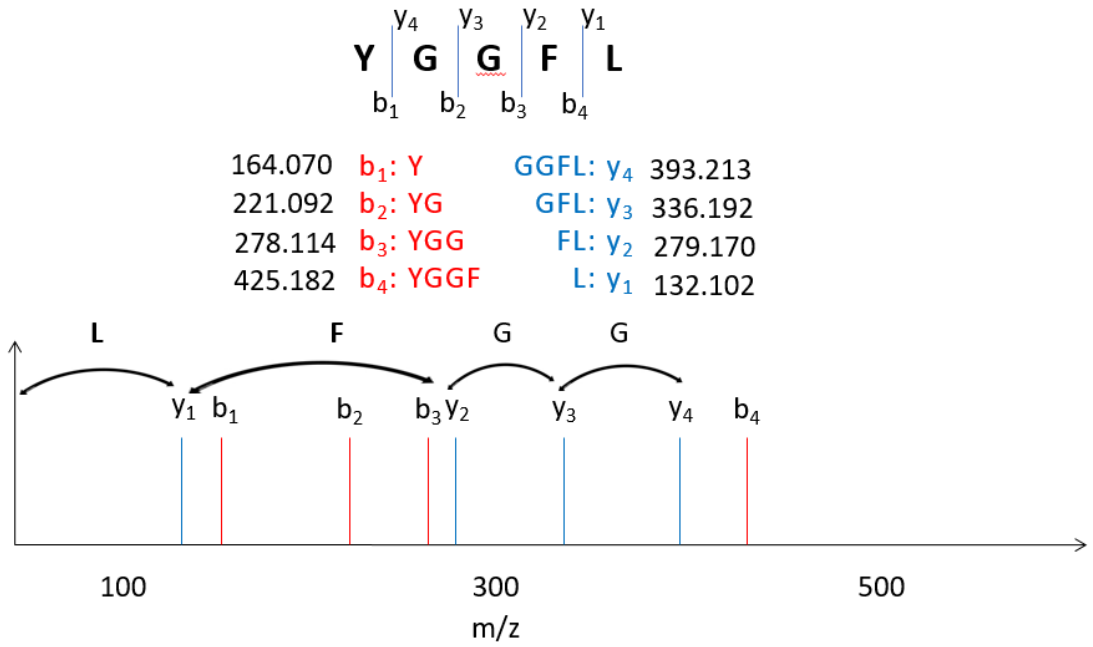

Рис. 3: Теоретическая фрагментация $b-$ и $y-$ионов пептида **YGGFL**. Фрагменты с массами приведены над спектрами. Двойные стрелки с подписями иллюстрируют факт того, что расстояние между двумя последующими фрагментарными ионами равно массе соответствующей аминокислоты.

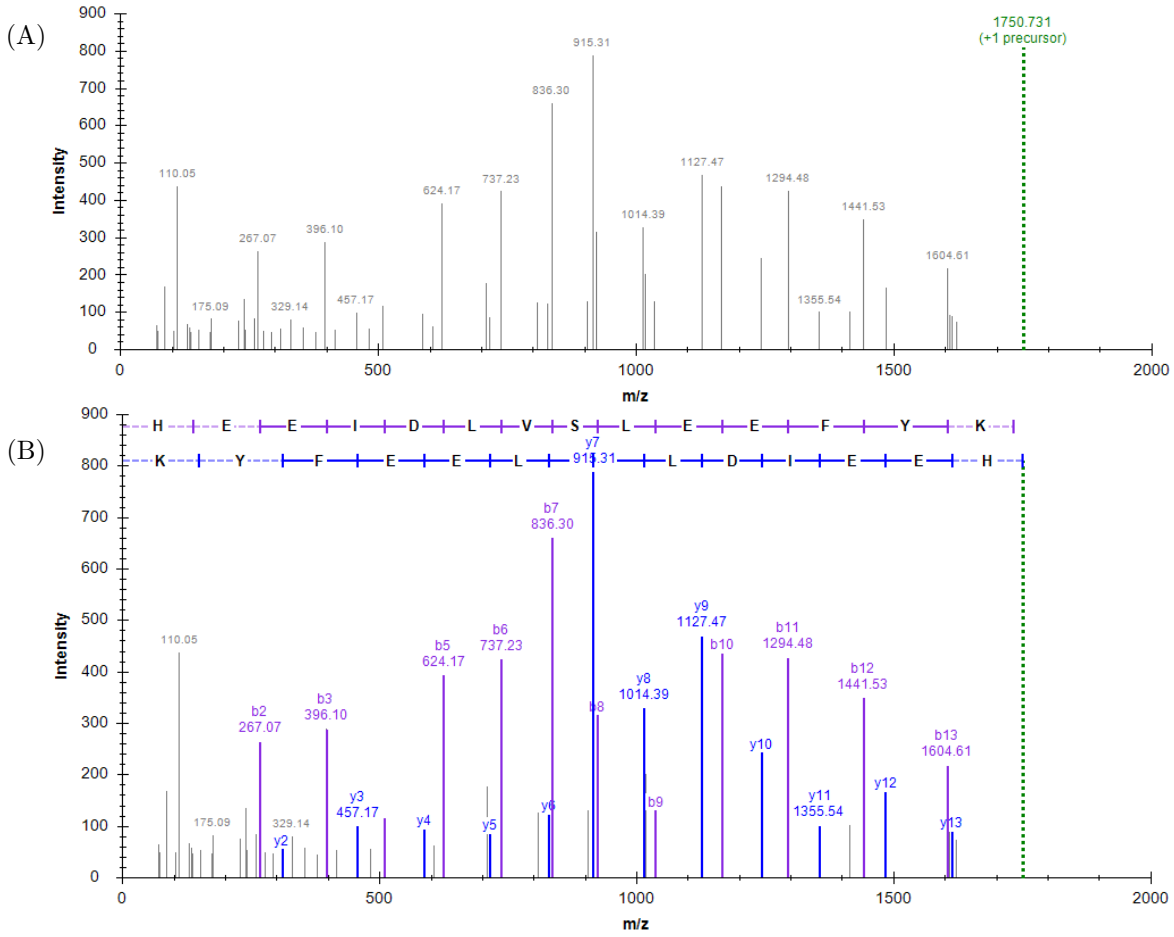

Рис. 4: Масс-спектр (A) для **HEEIDLVSLEEFYK** с возможной аннотацией (B).

Данный подход состоит из трех ключевых элементов: (1) база данных пептидов DB, (2) выбор биологически/химически вероятных пептидов, называемых пептидами-кандидатами ($CP(s_i) \subseteq DB$) для данного экспериментального спектра s_i, и (3) функция скоринга $\phi : S \times DB \rightarrow R$.

2.3.1 База данных пептидов

База данных пептидов получена из библиотек белковых последовательностей, таких как UniProt. Важно, что если биологические образцы получены от человека, то база данных пептидных данных должна быть построена из последовательностей белков человека.

Последовательности белка расщепляются в определенных местах в соответствии с правилами разрезания ферментом, используемым при обработке образца. Это также называется *in silico* расщепление белка [55]. На практике же возможны ситуации, когда ферменты не расщепляютс должным образом, или биологические образцы подвергаются некоторым модификациям. База данных пептидов должна учитывать подобные альтернативы. Стандартные варианты генерации пептидов с учетом означенных выше нетипичных ситуаций включают (см. Таблицу 1 для примера расщепления трипсином с возможными изменениями):

- Пропущенные разрезы: количество пропущенных мест расщепления. Этот параметр обычно находится в диапазоне 0-3.
- Неферментативный разрез: когда один или оба пептидных конца возникают в результате неферментативного расщепления.
- Посттрансляционные модификации (PTM): некоторые молекулы, например кислород или фосфор, могут присоединяться к аминокислоте и, следовательно, изменять массы фрагментарных ионов пептида (см. Рисунок 5).

Существует два основных типа посттрансляционной модификации: статический и переменный [52]. Статические модификации будут каждый раз применяться к определенной аминокислоте - например, карбамидометилирование (CAM) является результатом реакции остатков цистеина (C) с йодацетамидом [61]. Это приведет к изменению веса аминокислоты C на 57.02146 Da, а «статичность» означает, что она будет применяться ко всем цистеинам во всех пептидах. Переменная модификация теоретически может происходить в пептиде - и при расщеплении белков *in silico* может быть выбрано максимальное количество переменных модификаций на пептид. Примером переменной модификации может быть окисление метионина (которое приведет к увеличению массы метионина на 15.995 Da) или изобарический тандемный масс-тег (TMT) для лизина и N-концевых аминокислот (N-концевая модификация), приводящий к росту массы на 229.16293 Da.

2.3.2 Пептиды-кандидаты

Набор пептидов-кандидатов (CP) состоит из пептидов, у которых нейтральная масса родительского иона (MPA) равна нейтральной массе родительского иона s_i с точностью до

Оригинальный фрагмент белка
MEICRGLR|SHLITLLLFLFHSETICRPSGR|K|SSK|MQAFR|IWDVNQK|G...
Триптические пептиды
MEICRGLR, **MQAFR**, **IWDVNQK**, ...
Пропущенные расщепления = 1
IWDVNQKG, **SGRK**, ...
Неферментированные пептиды (случайные разрезы)
SHLITLLLF, **SGRKSS**, ...
Модифицированные пептиды
M(ox)EICRGLR, ...

Таблица 1: Фрагмент белка из *IPI: IPI00000045.1 | SWISS-PROT: P18510-1*. Вертикальными полосками отмечены места триптического расщепления. Место пропущенного расщепления окрашено (красным). Место, где трипсин не делает разрез из-за правила подавления пролином, выделено (фиолетовым). **(ox)** указывает на окисление метионина (M).

погрешности, специфичной для экспериментального прибора. Данное ограничение на кандидатов исключает из процедуры скоринга пептиды, молекулярные массы которых отличаются от массы родительского иона, так что они априори не могут быть правильными. В аннотации это ограничение ускоряет поиск и приводит к гораздо меньшему количеству ложных аннотаций. CP формально определяются как:

$$CP(s_i) = \{h_j : h_j \in DB, D(s_i, h_j, Z(s_i), \epsilon) \leq 0\} \tag{2}$$

где $Z(s_i)$ это состояние заряда родительского иона s_i, D это функция допустимого отклонения, ϵ является *коэффициентом допустимого отклонения массы* [54]. На практике функция допустимого отклонения обычно определяется одним из следующих способов:

- В Дальтонах:
  $D(s_i, h_j, Z(s_i), \epsilon) = \left| \frac{MPA(h_j)}{Z(s_i)} - MPA(s_i) \right| - \epsilon$
- В PPM (миллионных долях):
  $D(s_i, h_j, Z(s_i), \epsilon) = \left(MPA(s_i) - \frac{MPA(h_j)}{Z(s_i)} \right)^2 - \left(MPA(s_i) * \frac{\epsilon}{1000000} \right)^2.$

2.3.3 Функции скоринга

Функция скоринга измеряет своего рода сходство между экспериментальным и теоретическим спектрами, а более высокий показатель полученного в результате скоринга рейтинга указывает на лучшее соответствие спектра, основанное на подсчете экспериментальных и теоретических пиков в одной и той же позиции в спектрах s_i и h_j. Однако из-за погрешности измерения считается, что пики совпадают с точностью до небольшой погрешности для конкретного прибора δ. Например, если есть пик в позиции p m/z в экспериментальном спектре и пик в позиции q m/z в теоретическом спектре, то эти пики считаются совпадающими тогда и только тогда, когда $|p - q| < \delta$. Этот подход является более точным, хотя и вычислительно сложным.

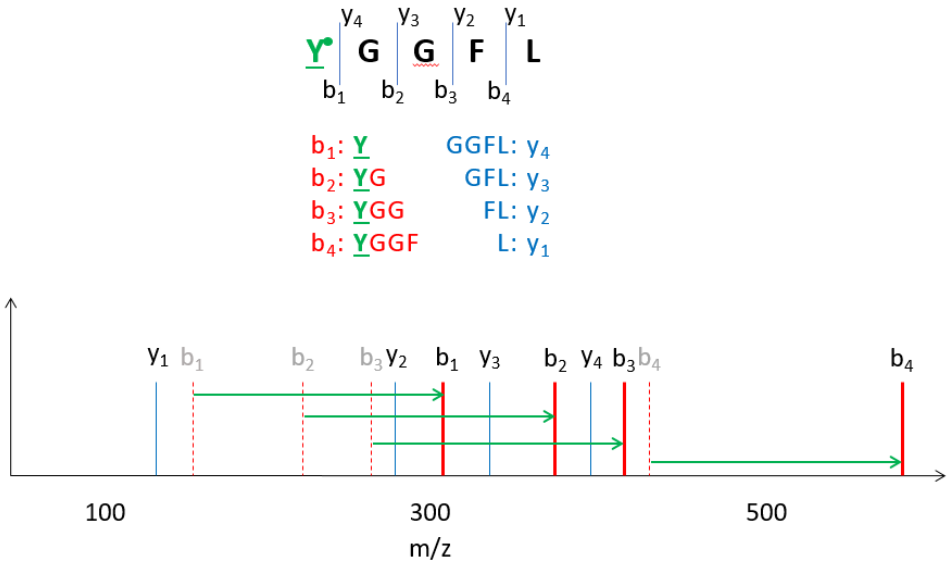

Рис. 5: Сульфатирование тирозина (Y) в **YGGFL** изменяет вес аминокислот и пептида, что приводит к сдвигу пиков в спектре. Одна PTM сдвигает ровно половину всех пиков в пептиде. Если модификация происходит не с первой аминокислотой, некоторые $y-$ионы также будут сдвинуты.

Другой широко используемый подход включает дискретизацию спектра, при которой каждый спектр преобразуется в вещественный вектор v, а пики помещаются в соответствующие позицию (ячейку, канал) вектора. Пик в позиции p m/z помещается в ячейку $k = \left[\frac{p+(Z-1)*MP}{Z*res} + 1.0 - offset\right]$ и значение $v[k]$ - интенсивность пика. Здесь MP обозначает массу протона. Если в одну и ту же ячейку попадает более одного пика, значение $v[k]$ обычно выбирается как максимум интенсивностей пиков, приходящихся на одину и ту же ячейку k. Отметим, что заряды пиков невозможно узнать в экспериментальных спектрах и для них $Z = 1$; однако пики в теоретических пептидах могут генерироваться с несколькими состояниями заряда, и Z может варьироваться. Для инструментов с низким разрешением обычно используется $res = 1.0005079, offset = 0.4$; для современных инструментов высокого разрешения: $res = 0.02, offset = 0.0$. Поэтому, если наибольшее возможное положение пика берется за 2000 m/z, то дискретизация приводит к 1999-мерному вектору для случая низкого разрешения данных и 100 000-мерному вектору для случая с высоким разрешением данных. Для теоретических спектров интенсивность пиков равна 1.0, что отражается в итоговых бинарных векторах. Недостаток этого подхода заключается в том, что пики, которые расположены ближе друг к другу, чем допустимая погрешность ($|p-q| < \delta$), могут оказаться в смежных векторных ячейках.

В оставшейся части этой диссертации любой спектр s_i или h_j обозначает дискретизированный вектор масс-спектра.

Стандартные функции скоринга:

- Shared Peak Count (SPC) определяемый как:

$$\mathrm{SPC}(s_i, h_j) = \sum_{k=1}^{N} \mathbb{I}(s_i[k] \neq 0) \times \mathbb{I}(h_j[k] \neq 0), \tag{3}$$

где N это число ячеек. Этот подход не учитывает интенсивности пиков.

- Inner product (IP):

$$\mathrm{IP}(s_i, h_j) = s_i^T h_j. \tag{4}$$

Эта функция учитывает интенсивности совпадающих пиков.

- HyperScore, представленный Fenyo et al. [17]:

$$\mathrm{HyperScore}(s_i, h_j) = \mathrm{IP}(q, t) \times N_b! \times N_y! \tag{5}$$

где $N_b!$ является факториалом числа совпадающих b-ионов, а $N_y!$ - то же самое, но для совпадающих y-ионов. Идея заключается в том, что, например, 4 совпадающих b-иона (или 4 совпадающих y-иона) могут указывать на лучшее совпадение экспериментального и теоретического спектров, чем совпадающие 2 b-иона и 2 y-иона.

- Функция кросс-корреляции (XCorr) была представлен вместе с SEQUEST [15] как

$$\mathrm{XCorr}(s_i, h_j) = \mathrm{IP}(s_i, h_j) - \frac{1}{151} \sum_{\tau=-75}^{+75} \mathrm{IP}(s_i, h_j[\tau])), \tag{6}$$

Первая часть функции определяет соответствие между экспериментальным и теоретическим спектрами, используя скалярное произведение соответствующих векторов. Вторая часть дает оценку среднего значения нулевого распределения из 151 случайного сопоставлений, полученную с помощью случайного теоретического пептида $h_j[\tau]$, сгенерированного смещением компонентов вектора h_j на τ позиций. Заметим, что теоретические спектры соответствуют реальным пептидным последовательностям, в то время как сдвиг его компонентов на $\pm\tau > 0$ позиций нарушает его семантику, и он не может быть связан с какой-либо реальной пептидной последовательностью исходной массы, что приводит к созданию случайного вектора. Следовательно, оценка XCorr показывает разность между количественной оценкой соответствия и оценочным средним значением нулевого распределения.

2.4 Валидация результатов поиска.

При поиске по базе данных каждому спектру присваивается один пептид с наибольшим результатом скоринга (см. Таблицу 2); однако это не означает, что данное соответствие является верным. Идентификация пептидов на основе поиска по базе данных (1) является неточной, что означает, что наблюдаемый спектр (1a) содержит много необъяснимых пиков, которые

происходят из-за неправильной фрагментации пептида или появления сторонних молекул, а также (1b) не содержит ожидаемых фрагментарных ионов, которые не наблюдаются в масс-спектрометре; помимо этого поиск по базе (2) является неполным, что означает, что база данных пептидов не является полной, то есть «правильный» пептид может быть не включен в нее [49]. На практике примерно 40%-80% аннотаций могут быть неверными. Следовательно,

Спектр		Лучший пептид-кандидат	Рейтинг
s_1	$\leftarrow$	h_{1_j}	0.39
s_2	$\leftarrow$	h_{2_j}	0.97
s_3	$\leftarrow$	h_{3_j}	0.42
s_4	$\leftarrow$	h_{4_j}	1.13
...	...	...	...
s_n	$\leftarrow$	h_{n_j}	0.01

Таблица 2: Список наилучших аннотаций спектров пептидами.

методы проверки правильности аннотации спектра имеют большое значение.

2.4.1 Фальшивые пептиды

Подход target-decoy (TDA) был введен в 2007 году [11] для проверки аннотаций спектра. Этот подход начинается с генерирования фиктивных пептидных последовательностей, называемых фальшивыми (*decoy*) пептидами, и в дальнейшем они объединяются с набором исходных реальных пептидов, называемых целевыми теоретическими (*target*) пептидами. Существует четыре распространенных способа создания фальшивых пептидов:

- protein reverse: в этом подходе вся белковая последовательность переставляется в обратном порядке, и фальшивые пептиды генерируются путем расщепления *in silico*, как в случае целевых теоретичсеких пептидов;
- protein shuffle: в этом случае аминокислоты всей последовательности белка перетасовываются, и при расщеплении *in silico* образуются фальшивые пептиды. Этот подход генерирует больше фальшивых пептидов, чем целевых, потому что многие белковые последовательности являются гомологами и, следовательно, имеют общие триптические пептиды. Схема protein shuffle будет перетасовывать все копии этих общих пептидов несколько раз, по одному разу для каждого вхождения в протеом. Результатом является база данных фальшивых пептидов, которая больше исходной базы данных целевых теоретических пептидов, что, в свою очередь, приводит к консервативной оценке FDR [23];
- peptide reverse: аминокислоты целевых теоретических пептидов переставляются в обратном порядке;
- peptide shuffle: аминокислоты целевых теоретических пептидов случайно перетасовываются.

Обычно концевые аминокислоты остаются на своих местах в процедурах генерации фальшивых пептидов на уровне изменения самих пептидов. Это необходимо, потому что концевые аминокислоты специфичны для фермента расщепления, и модификация этих концевых аминокислот приведет к потере информации о типе пептида, будь то целевой теоретический или фальшивый пептид. Например, для пептида ACCQPSTYK peptide reverse приводит к образованию AYTSPQCCK, а подход peptide shuffle может привести к ATQPCSCYK.

Спектр		Лучший пептид-кандидат	Рейтинг
s_1	$\leftarrow$	$h^*_{1_j}$	0.39
s_2	$\leftarrow$	h_{2_j}	0.97
s_3	$\leftarrow$	$h^*_{3_j}$	0.42
s_4	$\leftarrow$	h_{4_j}	1.13
$\dots$	$\dots$	$\dots$	$\dots$
s_n	$\leftarrow$	h_{n_j}	0.01

Таблица 3: Список лучших аннотаций спектров пептидами, фальшивые пептиды отмечены (*).

После генерации фальшивых пептидов и объединения со списком целевых теоретических пептидов выполняется стандартный этап поиска по базе данных (см. Таблицу 3). Затем ошибка в аннотации спектра может быть оценена на основе количества фальшивых пептидов, найденных в результате поиска при определенных предположениях, обсуждаемых в следующем разделе.

2.4.2 Оценка FDR

Предполагая, что более высокий рейтинг означает лучшую аннотацию, PSM можно упорядочить по их соответствующим рейтингам в порядке убывания. При любом конкретном пороговом значении PSM выше данного значения могут рассматриваться как правильные аннотации, также называемые принятыми PSM. На практике соотношение ложно-положительных результатов оценивается среди принятых PSM, где количество ложных срабатываний оценивается на основе количества PSM, содержащих фальшивые пептиды, называемых фальшивыми PSM. Процент ложных отклонений гипотезы (FDR) рассчитывается следующим образом:

$$\text{FDR} = E\left[\frac{FP}{R}\right], R > 0, \tag{7}$$

где FP указывает количество фальшивых PSM, а R обозначает количество целевых PSM, то есть PSM, которые содержат целевые теоретические пептиды, среди всех принятых PSM. В Levitsky et al. [44] было показано, что формула для расчета FDR в уравнении 7 приводит к смещению оценки, поэтому следует добавить $+1$ к FP перед делением на R.

Точный контроль над FDR требует следующих допущений:

A-1) Наборы целевых и фальшивых пептидов должны различаться,

A-2) Целевые и фальшивые пептиды должны генерироваться независимо,

A-3) Количество целевых и фальшивых пептидов должно быть примерно одинакового размера; в противном случае расчет FDR должен включать поправочный коэффициент на размер.

A-4) Любыой спектр должен быть отнесен к неправильному целевому или фальшивыми пептиду с равной вероятностью.

На практике пороговое значение для оценки выбирается таким образом, чтобы уровень FDR принятых PSM был на предварительно определенном уровне α.

2.4.3 Расчет q-значения

Q-значение PSM определяется как наименьший уровень α, при котором оно принимается на уровне α FDR. Например, если q-значение PSM равно 0.005, то оно принимается на уровне FDR 0.5 %, но не принимается, скажем, на уровне FDR 0,50001 %. Следует отметить, что q-значение PSM зависит не только от спектра и соответствующего ему набора пептидов-кандидатов, но также от аннотаций других экспериментальных масс-спектров.

Псевдокод вычисления q-значений показан в Алгоритме 1.

Algorithm 1: Расчет q-значений

Input : Список лучших PSMs: $\phi(s_1, h_{1_j}) \, , \ldots , \, \phi(s_n, h_{n_j})$
Output: $Q_1 \ldots Q_n$
(Сортировка по убыванию$(\phi(s_1, h_{1_j}) \, , \ldots , \, \phi(s_n, h_{n_j}))$)
$targets \leftarrow 0$
$decoys \leftarrow 1$
for $i = 1 \rightarrow n$ **do**
  if h_{i_j} целевой пептид **then**
    $targets \leftarrow targets \, + \, 1$
  else
    $decoys \leftarrow decoys \, + \, 1$
  end if
  $FDR_i \leftarrow \frac{decoys}{targets}$
  if $FDR_i \, > \, 1$ **then**
    $FDR_i \leftarrow 1$
  end if
end for
$Q_1 \leftarrow FDR_1, \, \ldots , \, Q_n \leftarrow FDR_n$
for $i = n - 1 \rightarrow 1$ **do**
  if $Q_{i+1} < Q_i$ **then**
    $Q_i \leftarrow Q_{i+1}$
  end if
end for

Пример расчета q-значений с использованием Алгоритма 1 приведен в Таблице 4.

Примеры кривых q-значений для различных поисковых методов представлены на Рисунке 6.

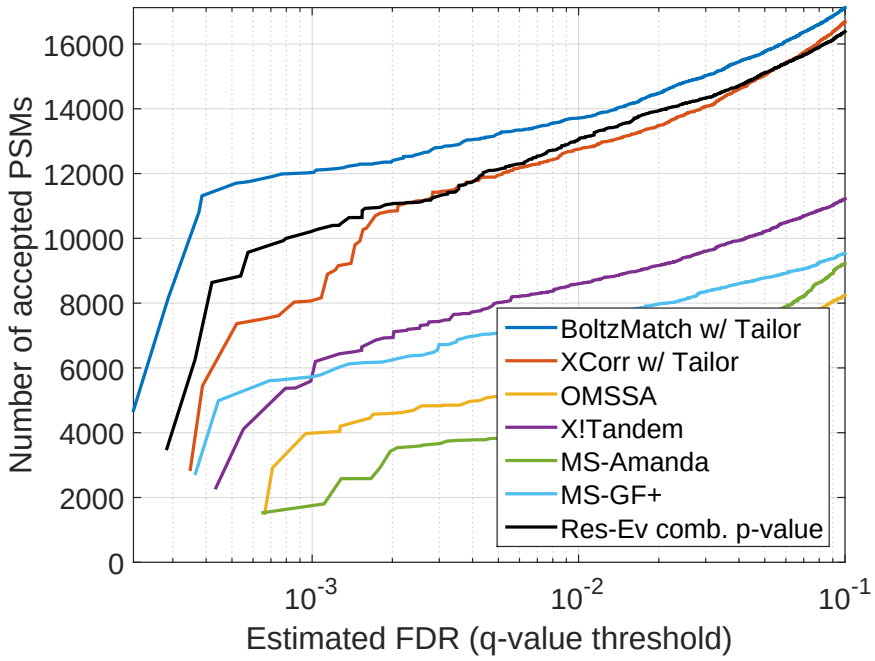

Рис. 6: Результаты аннотации спектра с использованием различных поисковых методов. На оси X отмечаются пороговые q-значения, а ось Y обозначает количество принятых PSM при заданном пороговом q-значении. Разные линии соответствуют разным методам; более высокие кривые указывают на более хороший результат аннотации.

Спектр	Лучший пептид-кандидат	Рейтинг	FDR	q-значение
s_6	h_{6_j}	2.03	$0/1 = 0$	$= 0$
s_9	h_{9_j}	1.96	$0/2 = 0$	$= 0$
s_{11}	$h^*_{11_j}$	1.54	$1/2 = 0.5$	$= 0.16$
s_4	h_{4_j}	1.13	$1/3 = 0.33$	$= 0.16$
s_7	h_{7_j}	1.11	$1/4 = 0.25$	$= 0.16$
s_2	h_{2_j}	0.97	$1/5 = 0.2$	$= 0.16$
s_{10}	h_{10_j}	0.65	$1/6 = 0.16$	$= 0.16$
s_3	$h^*_{3_j}$	0.42	$2/6 = 0.33$	$= 0.33$
s_1	$h^*_{1_j}$	0.39	$3/6 = 0.5$	$= 0.38$
s_5	h_{5_j}	0.28	$3/7 = 0.43$	$= 0.38$
s_8	h_{8_j}	0.25	$3/8 = 0.38$	$= 0.38$

Таблица 4: Илюютрация расчетов FDR и q-значений.

2.5 *De Novo* и гибридные методы идентификации пептидов

Принципиально другой подход к идентификации пептидов называется *de novo* секвенированием [46]. Этот подход используется для новых протеомов, когда базы данных последовательностей белков недоступны для новых видов (новых бактерий и т.д.). Преимущество секвенирования *de novo* заключается в том, что ему не требуется база данных белков-кандидатов, однако недостатком этого подхода является то, что статистические подходы не могут быть использованы для проверки идентификаций. Подходы, основанные на поиске по базе данных, используются, когда доступны базы данных последовательностей белков (например, для тканей человека). Методы *de novo* используют биологические и химические правила для определения подпоследовательностей в экспериментальном спектре s_i [4, 31].

Гибридные методы идентификации спектра объединяют метод *de novo* с методами поиска по базе данных, в которых подпоследовательности, идентифицированные подходом *de novo*, используются для фильтрации $CP(s_i)$ путем удаления теоретических пептидов h_j, которые не содержат любые подпоследовательности, определенные подходом *de novo*.

Результаты этой диссертации связаны с подходами, основанными на поиске по базе данных, и я не буду обсуждать методы *de novo* более подробно.

3 Свойства функций скоринга

Функции скоринга являются ключевыми в процессах идентификации пептидов. Хорошие функции скоринга должны быть:

a) *дискриминативными*, то есть они должны отличать правильные PSMs от неправильных,

b) *откалиброванными*, что означает наличие четко определенной и точной семантики,

c) *несмещенными*, то есть они с равной вероятностью присваивают каждый спектр неправильному целевому или ложному пептиду,

d) *универсальными*, то есть они хорошо работают со спектрами, сгенерированными с использованием различных конфигураций масс-спктрометрических приборов и экспериментальных протоколов. [23, 40].

3.1 Свойство дискриминативности

Хорошая дискриминативная способность функции скоринга означает, что распределение оценок, соответствующих правильным PSM, хорошо отделяемо от распределения неправильных PSM; следовательно, правильные PSM могут быть отделены от неправильных с помощью простых пороговых значений.

Функциям оценки в идентификации спектра препятствуют (а) наличие многих необъяснимых пиков, которые возникают из-за необычной фрагментации пептида или появления сторонних молекул, или (б) отсутствие ожидаемых фрагментарных ионов, которые не регистрируются в масс-спектрометре [49]. Функции скоринга пытаются смягчить негативные эффекты, вызванные этими проблемами (а), рассматривая вторичные продукты фрагментации ионов (SFIP), такие как ионы, полученные в результате потери молекул воды, монооксида углерода или аммиака, в дополнение к первичным фрагментарным ионам. Например, Andromeda [6] генерирует вспомогательные пики для продуктов потери воды или аммиака для теоретических пептидов, содержащих аминокислоты D, E, S, T или K, N, Q, R соответственно; в то же время популярная функция XCorr SEQUEST [15, 63] дополнительно включает в себя сигналы от соседних ячеек дискретизированного вектора спектра [13], SFIP, а также массы фрагментарных ионов с высоким зарядом в зависимости от состояния заряда родительского иона. Функция XCorr может быть формализована как:

$$\mathrm{XCorr}(s,h) = E(s,h) - Z(s,h) \tag{8}$$

для дискретизированного экспериментального s и теоретического h спектров, где E присваивает вес 50 совпадающим первичным фрагментарным ионам, обычно $b-$ и $y-$ионов, вес 25 совпадающим соседним пикам, и вес 10 совпадающим SFIP пикам, и $Z(s,h)$ является поправочным коэффициентом, который определяется как $Z(s,h) = \frac{1}{151}\sum_{\tau=-75}^{+75} E(s,h[\tau])$, где в $h[\tau]$

все элементы вектора h сдвинуты на τ шагов [15, 57]. Веса могут быть организованы в матрицу весов W, описываемую формулой $E(s,h) = s^T W h$, где T обозначает транспонирование.

Также были представлены несколько новых скоринговых функций и инструментов поиска по базам данных, включая Mascot [50], HyperScore из X!Tandem[17], Morpheus [60], и MS Amanda [10]; однако эти методы основаны на вручную созданных скоринговых функциях и дали в результате незначительные улучшения по сравнению с SEQUEST [38]. Недавние исследования были в основном сосредоточены на калибровке рейтингов, чтобы обеспечить четко определенную точную семантику, чтобы аннотации спектра можно было сравнивать друг с другом [35, 40, 36, 34, 57]; однако обсуждение дискриминирующей силы функций оценки часто игнорируется.

3.2 Свойство откалиброванности

Неоткалиброванные первичные количественные оценки могут давать различное качество сопоставления для разных спектров. Например, распределения PSM с наибольшими рейтингами для двух- и трехзарядных спектров на Рисунке 7 показывают, что первичный рейтинг 2.5 может означать корректную аннотацию для двухзарядных, но некорректную аннотацию трехзарядных молекул пептидов [35]. Методы калибровки оценки для конкретного спектра направлены на то, чтобы обеспечить своего рода нормализацию оценки, то есть чтобы рейтинги сопоставления теоретических пептидов для разных экспериментальных спектров были сравнимы между собой; таким образом можно выбрать один порог для принятия или отклонения аннотаций спектра. Калибровка позволяет получить гораздо больше аннотаций спектров при любом желаемом уровне FDR [35]. Методы калибровки рейтингов используют нулевое распределение и калибруют первичный рейтинг по среднему или хвосту нулевого распределения.

Стандартный подход калибровки рейтинга заключается в присвоении индивидуально для каждого спектра статистической значимости первичного рейтинга PSM путем оценки вероятности наблюдения случайного рейтинга, равного или превышающего наблюдаемый рейтинг PSM [24, 38, 39]. Такой метрикой является p-значение, которое на самом деле имеет четко определенную и точную семантику по различным экспериментальным протоколам и различным конфигурациям масс-спектрометрических инструментов. Успешность методов калибровки рейтинга зависит от того, насколько хорошо они аппроксимируют хвост или крайний хвост нулевого распределения для получения оценки p-значения. Некоторые методы используют аналитические модели, такие как биномиальное распределение в Andromeda [6] и MS Amanda [10], распределение Пуассона в Open Mass Spectrometry Search Algorithm (OMSSA) [20], распределение Вейбулла для XCorr [41], или распределение Гумбеля для Spectrum Specific P-value (SSPV) [56], и основаны на на предположении о том, что пики совпадают независимо между спектрами. Недостатки этих моделей заключаются в том, что (а) это предположение не является правдоподобным на практике [8], и (б) аналитические функции вероятности (PMF) биномиального или пуассоновского распределений не имеют функций распределения в замкнутых формах для вычисления p-значения моментально. В

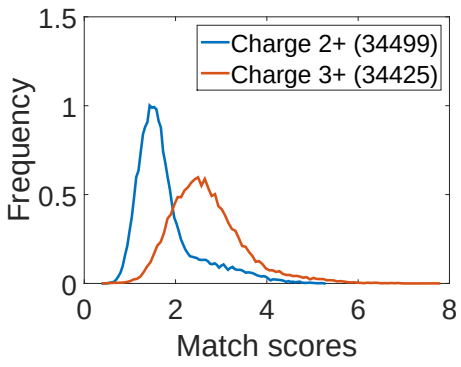

Рис. 7: Distributions of top scoring PSMs obtained with simple match on Yeast data for doubly (blue) and triply (red) charged precursor ions.

результате им требуется больше процессорного времени для суммирования по большему количеству PFM при гипотетических рейтингах PSM. Параметры экспоненциальных распределений (Вейбулла и Гумбеля) подбираются из эмпирических оценок PSM отдельно для каждого спектра.

X!Tandem [17] и Comet [14] подбирают подходящую модель линейной регрессии для оценки функции выживания нулевого распределения, чтобы откалибровать оценку для каждого экспериментального спектра. Comet использует log преобразование функции выживания, подбирает модель линейной регрессии, и вычисляет откалиброванную оценку, E-значение, путем экстраполяции модели линейной регрессии на наибольшую оценку рейтинга PSM. X!Tandem использует такой же подход - подбирает модель линейной регрессии к эмпирической функции выживания логарифма HyperScores [17]. Оба подхода предполагают, что хвосты нулевого распределения убывают экспоненциально; однако это предположение не подвергалось критическому анализу.

Недостатки методов калибровки оценок, основанных на подборе конкретных параметрических моделей, заключаются в том, что они не могут быть напрямую обобщены для других функций скоринга, и что параметрическое распределение, параметры которого оцениваются с использованием общих распределений оценок PSM, может быть неточным в крайнем хвосте [56].

Другие типы методов оценки p-значений используют точное нулевое распределение, полученное при подсчете всех возможных пептидных последовательностей, которые имеют же самые — с точностью до специфической погрешности прибора — массы родительского иона, как и у наблюдаемого спектра [38, 39, 40, 30, 45]. Точное перечисление всех последовательностей с вычислительной точки зрения невозможно; поэтому используются методы

динамического программирования для подсчета количества пептидов с определнным значением рейтинга в нулевом распределении. Эти методы действительно приводят к отличной калибровке рейтингов; однако у них есть несколько недостатков:

1. Они требуют правильной оценки частот аминокислот в базе данных пептидов.
2. Расчет элементов таблицы динамического программирования требует значительного количества процессорного времени.
3. Подход динамического программирования требует, чтобы функция оценки была аддитивной [30].
4. Метод динамического программирования не работает для основанных на сопоставлении пиков функций скоринга (например, XCorr), используемых на данных с высоким разрешением точности определения массы фрагментов, поскольку фрагментарные ионы аппроксимируются суммой дискретизированных масс аминокислот в методе динамического программирования, которая, в свою очередь, может отличаться от дискретизированной массы всего фрагментарного иона в настройках высокого разрешения. Следует отметить, что для данных MS2 с низким разрешением потеря информации из-за дискретизации практически не создает проблем на практике. Это подробно обсуждается Lin et al. [45].

Чтобы преодолеть многие из проблем, упомянутых выше, эмпирические p-значения PSM могут быть оценены с помощью скоринга спектров по большому количеству, скажем, 10K, фальшивых пептидов [35]. В этом сценарии можно получить хорошо откалиброванные p-значения для любого типа функции скоринга, на данных MS2 с высоким или низким разрешением - хотя и за счет большего использования процессорного времени. Рисунок 8 иллюстрирует и сравнивает принципы методов калибровки рейтингов по нулевому распределению.

3.3 Свойство несмещенности

Чтобы получить точный контроль и оценку FDR, неправильные аннотации спектра должны присваиваться как для целевых, так и для фальшивых пептидов с равной вероятностью. Стандартные функции скоринга отвечают этому условию, потому что они не способны различать целевые и фальшивые пептиды. Однако методы, основанные на машинном обучении, в которых участвуют целевые и фальшивые пептиды в обучении для повышения точности аннотации спектра, могут отдавать предпочтение пщелевым пептидам или аннотациям, содержащим целевые пептиды. Это приводит к смещенной оценке FDR.

3.4 Свойство универсальности

Различные приборы, протоколы экспериментов и параметры поиска по базе данных оказывают влияние на наблюдаемые экспериментальные спектры. Например, тип ионизации может различаться в разных приборах, и в результате экспериментальные спектры могут

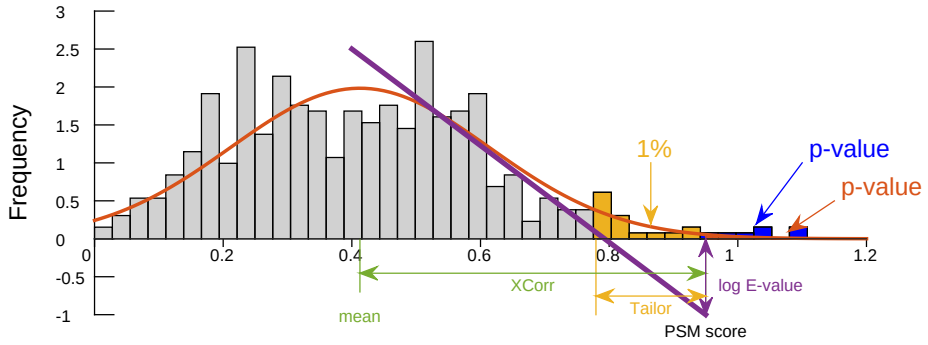

Рис. 8: Иллюстрация принципов подходов калибровки рейтинга PSM для нулевого распределения, обозначенного серым. Нулевое распределение было получено в результате оценки реального спектра. (Зеленый) XCorr калибрует рейтинг соответствия спектров, измеряя разницу между оценкой PSM и приближением среднего значения случайных оценок соответствия (Comet, Sequest, Tide). (Пурпурный) Методы, основанные на регрессии, подбирают наиболее подходящую линию на эмпирической функции выживания на основе гистограммы случайных оценок и экстраполируют E-значение в тех случаях, когда оценка PSM попадает на линию регрессии (Comet, X! Tandem). (Синий) Эмпирические p-значения вычисляются из точного нулевого распределения, полученного методами динамического программирования (XCorr exact p-value в Tide, MS-GF+) или методами Монте-Карло [35]. (Красный) P-значения рассчитываются с использованием аналитических функций плотности вероятности (OMSSA, Andromeda, Morpheus, SSPV, Weibull calibration of XCorrs). (Желтый) Метод Tailor калибрует рейтинги по верхнему 100-му квантилю распределения, то есть относительно рейтинга, который имеет p-значение 0.01.

иметь разные распределения пиков. Кроме того, экспериментальные протоколы, возможность модификаций или пропущенных расщеплений, влияют на набор теоретических пиков. Метод машинного обучения, обученный для набора данных определенного типа, может не обязательно обобщаться на другие спектры, генерируемые с помощью различных типов инструментов и экспериментальных протоколов. Например, признаки, извлеченные из данных спектров, полученных при фрагментации методом высокоэнергетической столкновительной диссоциации (HCD), могут не подходить для данных, полученных при фрагментации на основе столкновительной диссоциации (CID) или диссоциации с переносом электронов (ETD).

3.5 Изучение новых функций скоринга

Особенность обучения функций скоринга для аннотации спектров заключается в том, что невозможно получить заранее аннотированный набор масс-спектров, потому что наблюдатели не могут зайти внутрь масс-спектрометра, визуально наблюдать за молекулами и фиксировать спектры, которые они производят. Поэтому в области анализа данных масс-спектрометрии обучение с учителем вместе со сценарием разбиения выборки на обучающую, валидационную и тестовую не применяется; вместо этого используются методы машинного обучения (ML) с частичным привлечением учителя, и при таком подходе небольшое количество данных аннотируется стандартными методами поиска по базе данных и используются в качестве обучающих данных. Поскольку частичное привлечение учителя возможно без вовлечения человека в цикл, это можно сделать для всего объема аннотируемых данных.

Следовательно, вопрос заключается в том, насколько хорошо метод может быть обобщен для получения большего количества аннотаций на одном и том же наборе данных по сравнению со стандартной или другими функциями скоринга. Описанный подход к обучению был представлен в этой области в 2007 году вместе с инструментом Percolator[32], и с тех пор стал стандартным для распознавания данных масс-спектрометрии; хотя он и был представлен как метод с частичным привлечением учителя.

К сожалению, при таком подходе существует потенциальная опасность. Метод ML может научиться отдавать предпочтение целевым пептидам, то есть спектр, сопоставленный с целевым пептидом, может давать систематически более высокое значение рейтинга, чем когда он сопоставлен с фальшивыми пептидами. Это может привести к смещенной оценке FDR без каких-либо признаков проблем обучения модели. Поэтому важно показать, что улучшение аннотации спектра, сделанное функцией скоринга на основе машинного обучения, не возникает из-за этого смещения.

4 Резюме диссертационных статей

Основным результатом этой диссертации является метод BoltzMatch, который представляет собой функцию скоринга для аннотации спектра на основе стохастической нейронной сети. BoltzMatch был задуман как функция скоринга с высокой дискриминирующей способностью. BoltzMatch моделирует совместную вероятность наблюдения экспериментального s и теоретического h спектров, смоделированных с помощью ограниченной машины Больцмана (RBM), и она определяется как

$$p(s,h) = \frac{1}{Z}\exp\{E(s,h)\}, \tag{9}$$

где теоретический спектр h трактуется как ненаблюдаемая скрытая переменная, идеализированная версия наблюдаемого экспериментального спектра s, который содержит необъяснимые пики и неполные серии фрагментарные ионы. $E(s,h) = s^T W h$ называется энергетической функцией, а Z - нормировочным фактором, который определяется как $Z = \sum_{s',h'}\exp\{E(s',h')\}$ для всех возможных векторов s', h' и для которых параметры в W должны быть получены из наблюдаемых данных масс-спектрометрии. Логарифмическая вероятность $\log p(s,h) = E(s,z) - \log Z$ напоминает функцию XCorr, определенную в уравнении 8. С одной стороны, можно грубо рассматривать XCorr как логарифм правдоподобия RBM, настроенный вручную, а с другой стороны, можно грубо рассматривать BoltzMatch как обобщение XCorr, в котором признаки извлекаются из данных. Рисунок 9B иллюстрирует сопоставление экспериментального и теоретического спектров методами XCorr и BoltzMatch.

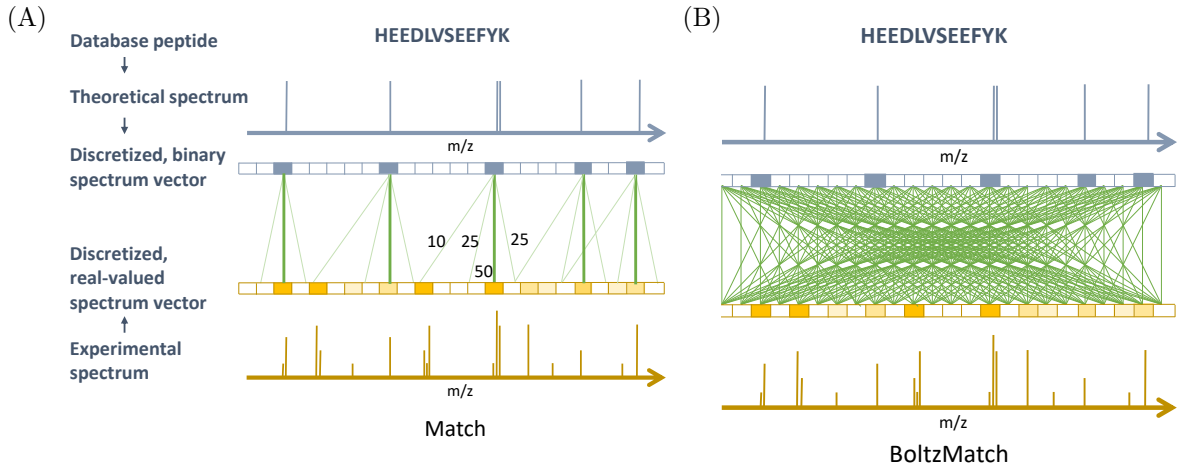

Рис. 9: Графические модели функций оценки XCorr и BoltzMatch. (A) XCorr присваивает совпадающим ионам вес 50, соседним пикам вес 25 и пикам на расстоянии потерь молекул вес 10; значения весов были настроены вручную. (B) Полностью связанная стохастическая нейронная сеть, BoltzMatch, для сопоставления наблюдаемых спекторв с теоретическими. BoltzMatch рассматривает связь между всеми парами пиков и учится взвешивать их исключительно на основе данных масс-спектрометрии.

4.1 Обучение BoltzMatch

Обучение RBMs проводится с помошью оценки максимального правдоподобия

$$\tilde{w} = \underset{w}{\operatorname{argmax}} \log p_w(s) \tag{10}$$

где s обозначает экспериментальный спектр, а p_w обозначает распределение Гиббса, параметризованное с помощью w и моделируемое ограниченной машиной Больцмана. Веса обновляются путем вычисления производных $\log p_w(s)$ по параметрам модели, то есть,

$$w_{i,j}^{(t+1)} = w_{i,j}^{(t)} + \frac{\partial \log p_w(s)}{\partial w_{i,j}}, \tag{11}$$

где t указывает на номер шага итерации. Взятие производных приводит к уравнению

$$\frac{\partial \log p_w(s)}{\partial w_{i,j}} = \sum_{h'} p_w(h' \mid s)\ s[i]\ h'[j]\ - \sum_{s',h'} p_w(s',h')\ s'[j]\ h'[i], \tag{12}$$

где первое суммирование проходит по всем возможным бинарным векторам h', а второе суммирование идет по всем возможным парам векторов s' и h'. Условная вероятность $p_w(h\ mids)$ определяется как $p_w(h = 1 \mid s) = \prod_i \sigma(\sum_{j=1}^{n} w_{ij} v_j)$, где $\sigma(a) = (1 + \exp(-a))^{-1}$ - функция сигмоида [18]. Обучение RBM, как известно, является сложной процедурой (а), когда задействованы латентные переменные, и (б) потому, что для выборки коэффициента нормализации Z [27] используются методы Монте-Карло по схеме марковской цепи (MCMC), чтобы избежать полный перебор s' и h'. Чтобы сделать обучение BoltzMatch более эффективным, мы разработали несколько приемов для решения этих проблем, используя особенности данных масс-спектрометрии:

1. Наша модель ограничивается только наблюдаемыми спектрами s' и возможными теоретическими спектрами h', которые кодируют реальные пептиды. Кроме того, мы определяем $p_w(h', s') = 0$ всякий раз, когда массы родительских ионов этих спектров не совпадают с точностью до погрешности для конкретного прибора. Обратите внимание, что $p_w(h', s') = 0$ может привести к проблемам при логарифмировании, поэтому мы просто просто избегаем рассмотрения таких пар спектров на практике. Обратите внимание, что мы считаем это предположение правдоподобным при идентификации спектра для поиска по базе данных.

2. Для каждого экспериментального спектра рассматривается только один теоретический пептид h, который может быть ответственным за генерацию наблюдаемого спектра s; следовательно, мы ожидаем $p_w(s, h) >> 0$, в то время как мы ожидаем $p_w(s, h') \approx 0$ для всех других теоретических пептидов h' в пределах допустимого отклонения массы.

Это приведет к упрощению уравнения 12 в следующем виде:

$$\frac{\partial \log p_w(s)}{\partial w_{i,j}} \approx \frac{\partial \log p_w(s,h)}{\partial w_{i,j}} = p_w(h \mid s)\ s[i]\ h[j]\ - \sum_{s',h'} p_w(s',h')\ s'[j]\ h'[i], \tag{13}$$

где h - теоретический пептид, ответственный за генерацию наблюдаемого спектра s. К сожалению, правильный теоретический пептид не известен. Поэтому выполняется стандартный шаг поиска по базе данных, чтобы определить (возможно) правильный теоретический спектр для каждого экспериментального спектра с q-значением менее 0.005 перед обучением BoltzMatch.

3. Второе слагаемое уравнения 12 включает в себя перечисление всех возможных векторов h'; однако большинство из них не соответствуют биологически правдоподобным векторным представлениям каких-либо пептидов. Например, рассмотрим вектор h, в котором каждая вторая ячейка заполнена единицей, а все остальные заполнены нулями; такие векторы могут быть исключены из перечисления. Поэтому мы ограничиваем второе слагаемое пептидами-кандидатами наблюдаемого спектра s, что приводит к следующей формуле:

$$\frac{\partial \log p_w(s,h)}{\partial w_{i,j}} \approx p_w(h \mid s)\ s[i]\ h[j]\ - \sum_{h' \in CP(s)} p_w(s,h')\ s[j]\ h'[i], \tag{14}$$

где $CP(s)$ обозначает набор пептидов-кандидатов экспериментального спектра s. Обратите внимание, что при стандартном обучении RBM второе слагаемое аппроксимируется с использованием методов MCMC; однако, по нашему мнению, подобное сэмплирование вряд ли приведет к какому-либо биологически правдоподобному вектору, который в этом случае может быть связан с любой реальной пептидной молекулой.

4. Наблюдаемый набор данных спектра может содержать "вездесущие"пики, которые появляются почти в каждом спектре в одном и том же месте m/z. Например, образцы в наборе данных HumVar были подготовлены с использованием шестиплексного изобарического тандемного масс-тега TMT, который имеет ассоциированный вес 229.16293 Da, и следовательно можно наблюдать пики в районе 230 m/z и 115 m/z почти во всех экспериментальные спектры. Эти пики, возможно, соответствуют однозарядным и двухзарядным остаткам масс-тега TMT. Эти вездесущие пики не содержат полезной информации для идентификации спектра, но они вмешиваются в генеративное моделирование, поскольку они могут коррелировать со всеми другими пиками. Чтобы смягчить влияние этих вездесущих пиков, мы добавили диверсифицирующую регуляризацию [58] −, подробно обсуждаемую в следующем подразделе −, в следующем виде:

$$DR = \sum_{s_i, s_j \in MB} h_i^T h_j, \tag{15}$$

к целевой функции, определенной в уравнении 14, где s_i, s_j - наблюдаемые пары спектра

из данной мини-группы MB и $h_i \sim p(h_i \mid s_i) = \sigma(s^T W)$ (h_j определяется аналогично), где $\sigma(a) = (1 + \exp(-a))^{-1}$ - функция сигмоиды.

Регуляризованное обучение BoltzMatch проводилось путем оптимизации $\log p_w(s, h)$ посредством оценки максимального правдоподобия:

$$\tilde{w} = \text{argmax}_w \left\{ \sum_{(s,h) \in D} \log \left(\frac{\exp(E_w(s,h))}{Z_s} \right) + \alpha DR \right\}, \tag{16}$$

где $Z_s = \sum_{h \in CP(s)} \exp E_w(s,h)$ и $(s,h) \in D$ обозначает PSM, имеющие q-значения менее 0.005, которые были получены при стандартном поиске по базе данных и α обозначает параметр регуляризации.

4.2 Диверсифицирующая регуляризация

Диверсифицирующая регуляризация (DR) была введена в качестве общего метода регуляризации, чтобы помочь обучить произвольные глубокие генеративные и дискриминативные модели. Метод DR был опубликован в виде отдельной статьи [58]. Здесь я привожу краткое описание метода и результатов.

Глубокие модели [3], особенно глубокие нейронные сети (DNN), итеративно обрабатывают данные на различных уровнях абстракции как $\mathbf{s} = \mathbf{h}_0 \rightarrow \mathbf{h}_1 \rightarrow \mathbf{h}_2 \rightarrow \cdots \rightarrow \mathbf{h}_L = y$. Первый слой $\mathbf{s}$ является слоем необработанных данных, а более глубокие слои $\mathbf{h}_l$ стремятся обеспечить более высокую абстракцию данных, часто называемую признаками. Последний уровень может соответствовать либо меткам класса y в задачах классификации, либо некоторой другой высокоуровневой сущности в задачах генерации. В каждом слое используется монотонная, нелинейная, так называемая функция активации $g_l(\mathbf{h}_l^T \theta_l) \rightarrow \mathbf{h}_{l+1}$ для преобразования объектов $\mathbf{h}_l$ в $\mathbf{h}_{l+1}$, где θ_l обозначает параметризацию преобразования объекта на данном слое.

Для набора данных $\mathcal{D} = \{\mathbf{s}_i\}$ параметры модели оцениваются путем максимизации вероятности правдоподобия данных, которая определяется как:

$$l(\theta; \mathcal{D}) = \log P(\mathcal{D}; \theta) = \sum_{\mathbf{s}_i \in \mathcal{D}} \log \sum_{\mathbf{h}} P(\mathbf{s}_i, \mathbf{h}; \theta). \tag{17}$$

Нижняя граница уравнения 17 может быть записана в виде:

$$\mathcal{L}(\theta, \phi; \mathcal{D}) = \sum_{\mathbf{s} \in \mathcal{D}} \log P(\mathbf{s}; \theta) - \sum_{\mathbf{s} \in \mathcal{D}} KL\left(Q_{\mathbf{s}}^{\phi},\ P_{\mathbf{s}}^{\theta}\right). \tag{18}$$

Это означает, что жесткая нижняя граница может быть достигнута путем минимизации расхождения Кульбака-Лейблера (KL) между вариационным распределением Q и точным апостериорным распределением P.

Обучение θ параметров в глубоких моделях является сложным, и сам процесс часто рассматривается скорее как искусство, а не наука. Существует четыре основных проблемы с

обучением глубоких моделей для задач классификации:

I. Обучение глубоких генеративных моделей с помощью самообучающегося послойного подхода не использует метки классов, поэтому важная информация может быть упущена.

II. Когда изучается генеративная модель, трудно отслеживать обучение, особенно на более высоких уровнях [22]. Для DNN метод обратного распространения страдает от проблемы, известной как затухающие градиенты [16].

III. В принципе, генеративная модель может быть произвольно приспособлена к данным [59, 29, 26], на практике процедура оптимизации со скрытыми переменными может застревать в локальных минимумах.

IV. Структура модели часто указывается заранее, и разработанная модель может не соответствовать данным. В частности, количество скрытых нейронов или слоев часто определяется интуитивно или согласно привычкам экспериментатора; однако трудно дать точную оценку количества скрытых компонентов.

В качестве решения вышеупомянутых проблем был введен новый метод регуляризации, называемый диверсифицирующей регуляризацией (DR), для скрытых модулей для обучения глубоких моделей для задач классификации. В принципе, предлагаемый регуляризатор поддерживает разное абстрактное представление для двух выборок данных, принадлежащих разным классам. Эта регуляризация обозначается как $D(\mathbf{h}_p^{(l)}, \mathbf{h}_q^{(l)})$, где $\mathbf{h}_p^{(l)}$ и $\mathbf{h}_q^{(l)}$ являются абстрактными представлениями данных $\mathbf{s}_p$ и $\mathbf{s}_q$ (соответственно) на слое l, и данные имеют различные типы ($y_p \neq y_q$). Для оценки максимального правдоподобия генеративных моделей DR определяется в терминах функции расхождения $D(Q_{s_p}, Q_{s_q})$ и включается в уравнение 18 как дополнительное слагаемое. Для дискриминационного обучения DNN с использованием обратного распространения DR определяется как функция расстояния и включается в целевую функцию в качестве аддитивной функции потерь.

В качестве примера практического применения, слои в сетях глубокого убеждения (DBN) могут быть предварительно предобучены с помощью ограниченных машин Больцмана (RBM).

Диверсифицирующая регуляризация для вариационной оптимизации вводится для отдельных факторов Q_s^j следующим образом:

$$D_H(Q^{\phi}_{\mathbf{s}_p}, Q^{\phi}_{\mathbf{s}_q}) = 1 - \sum_{h=\{0,1\}} \sqrt{Q^{j,\phi}_{\mathbf{s}_p}(h) Q^{j,\phi}_{\mathbf{s}_q}(h)} \tag{19}$$

Мы протестировали влияние DR на обучение сети глубоких убеждений (DBN) [43] на наборе данных CIFAR-10 [42]. Набор данных содержит маленькие (32×32) цветные картинки 10 классов, 50000 экземпляров для обучения и 10000 экземпляров для тестирования. Мы создали DBN из 11 слоев, каждый из которых имеет 500, 300, 200, 150, 100, 80, 60, 50, 30, 20 скрытых и 10 выходных блоков соответственно. Веса в каждом слое были предварительно обучены RBM с DR и без DR в качестве базового решения (контроля).

Результаты показаны на Рисунке 10. Первые 10 графиков показывают логарифмы псевдоправдободобия, полученные во время обучения RBM с DR (сплошной) и без DR (пунктир) на каждом уровне. Последние два графика в последнем ряду показывают величину функции потерь и ошибку на тесте во время тонкой настройки, когда она применялась после регуляризированного (сплошной) и нерегуляризированного (пунктирной) предобучения. Эти графики показывают, что точная настройка достигла гораздо более быстрой сходимости и гораздо лучших обобщающих результатов, когда веса были предварительно обучены с использованием DR. Эти результаты показывают, что DR помогает при обучении достичь лучшего локального минимума и меньшей ошибки обобщения.

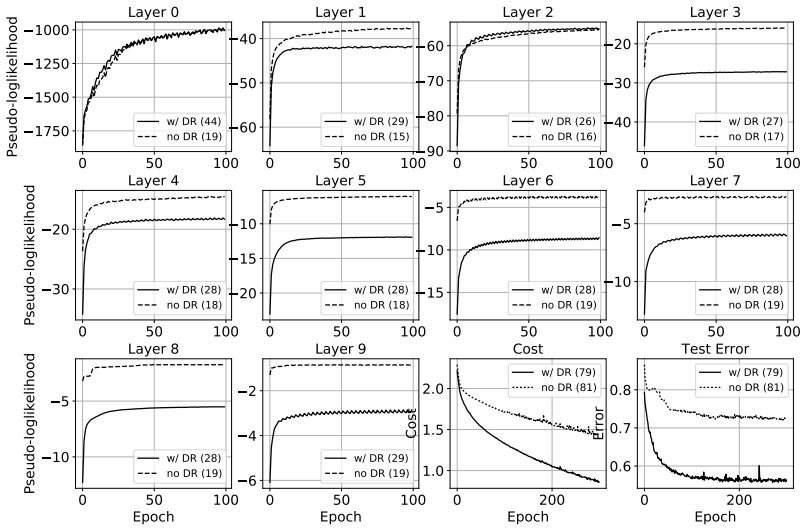

Рис. 10: Кривые обучения во время обучения RBM и DBN. Числа в скобках в описаниях указывают время выполнения в минутах. Функция потерь определяется как перекрестная энтропия.

В целом, улучшение с помошью DR в случае DBN может быть объяснено тем, что DR включает в себя информацию о классах, что может не максимизировать напрямую логарифм правдоподобия, но помогает в задачах, где разные типы данных имеют разные абстрактные представления. С другой стороны, DR позволяет извлечь пользу из дискриминативного обучения глубоких моделей, поскольку оно может обеспечить хорошие градиенты на ранних слоях, что напрямую помогает справиться с проблемой затухающего градиента.

4.3 Валидация BoltzMatch в идентификации спектра

После обучения BoltzMatch рейтинги PSM были рассчитаны как $\log p(s,h) = E(s,h) - log(Z_s)$, где $Z_s = \sum_{h' \in CP(s)} E(s,h')$ и экспериментальный спектр аннотируется теоретическим пептидом $\tilde{h}$, который дает наивысшую рейтинг $\tilde{h} = \text{argmax}_{h' \in CP(s)} \log p(s,h')$. Эти оценки не откалиброваны, и правильные методы калибровки могут привести к увеличению числа аннотаций спектра [35, 57]. Чтобы откалибровать показатель BoltzMatch с помощью метода XPV, пусть $PV_s(c)$ обозначает p-значение рейтинга c для данного спектра s, рассчитанного с помощью метода XPV. Тогда p-значение спектра s и рейтинг $c = p(s,h)$, полученные с

помощью BoltzMatch, могут быть определены как

$$PV_s(c) = PV_s(p(s,h)) \tag{20}$$
$$= PV_s(\log p(s,h)) \tag{21}$$
$$= PV_s(E(s,h) - \log Z_s) \tag{22}$$
$$= PV_s(E(s,h)) = PV_s((s^T W)h) \tag{23}$$
$$= PV_s(s_{BM}h), \tag{24}$$

где уравнение 21 следует из того факта, что функция log выполняет монотонное преобразование, которое не влияет на p-значение любых распределений, а уравнение 22 следует из того факта, что коэффициент нормализации $\log Z_s$ является константой, зависящей от спектра, и ее можно опустить, $s_{BM} = s^T W$, и $s_{BM}h$ обозначает скалярное произведение двух векторов s_{BM} и h. Поэтому рейтинги BoltzMatch можно калибровать с помощью любых стандартных методов калибровки рейтингов XPV с использованием преобразованных экспериментальных спектров s_{BM}.

Однако методы XPV не работают с данными высокого разрешения MS2, данный факт обсуждался обсуждался в Разделе 3.2 в Пункте 4. Мы разработали новый метод калибровки рейтингов, названный Tailor, который хорошо работает не только с BoltzMatch, но и с любыми функциями скоринга на данных высокого и низкого разрешении, и Tailor не основывается на каких-либо предположениях о законе распределения рейтингов. Этот метод будет подробно описан в следующем разделе.

4.4 Калибровка рейтингов PSM методом Tailor

Подход Tailor - это непараметрический, эвристический метод калибровки рейтингов PSM, который калибрует результат скоринга PSM путем деления рейтингов на верхний 100-ый квантиль эмпирических, специфичных для спектра нулевых распределений (т.е. результат скоринга с ассоциированным p-значением 0.01 в хвосте распределения), наблюдаемый при поиске по базе данных. Давайте рассмотрим экспериментальный спектр e, который соответствует N различным последовательностям пептидов-кандидатов на этапе поиска по базе данных, что дает следующие положительные рейтинги PSM: $s_1, s_2, \ldots, s_N > 0$. Предположим пока, что N достаточно велико и что эти оценки отсортированы в порядке убывания; таким образом, экспериментальный спектр e должен быть аннотирован пептидной последовательностью, которая дает оценку s_1. Эти оценки составляют основу эмпирического нулевого распределения для спектра e. 100 квантилей определяются 99 точками разреза, разделяющими диапазон распределения вероятностей на 100 непрерывных интервалов с равными вероятностями. Последняя (99-я) оценка 100 квантилей эмпирического нулевого распределения, обозначенная Q100, получается путем выбора оценки PSM в позиции $i^* = [N/100]$, где [.] обозначает стандартную операцию округления. Следовательно, $Q100 = s_{i^*}$, а метод Tailor

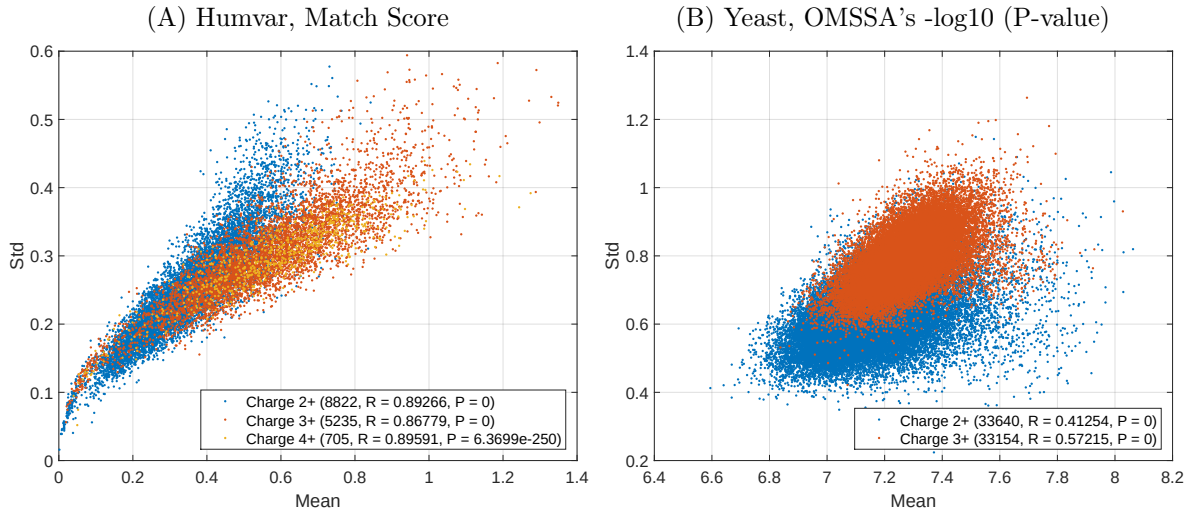

Рис. 11: Корреляция между средним значением и стандартным отклонением (std) эмпирических нулевых распределений. Каждая точка представляет среднее и std рейтингов соответствий пептидов-кандидатов одного экспериментального спектра. Оценка была выполнена с использованием набора данных HumVar с прямым сопоставлением Match score (A) и Yeast с OMSSA - логарифм p-значения (B) с набором фальшивых пептидов. Цвета указывают состояния заряда; числа в скобках показывают количество спектров, коэффициенты корреляции (R) и p-значения (P) для проверки гипотезы об отсутствии связи между средним и std (нулевая гипотеза).

калибрует первичные рейтинги сопоставлений по формуле

$$\tilde{s}_i = \frac{s_i}{Q100} \tag{25}$$

для $i = 1, \ldots, N$, которые являются различными рейтингами Tailor.

Метод Tailor использует хвост наблюдаемого нулевого распределения, который может быть неточным, однако случайные оценки наблюдаются на этапе поиска по базе данных, но не крайний хвост, где выборки редки. Это противоречит методам оценки точного p-значения (XPV, MSGF +), которые перечисляют все случайные оценки, в том числе в крайнем хвосте за счет использования процессорного времени, чтобы получить точное эмпирическое нулевое распределение. Поэтому Tailor быстр, хотя и менее точен, а точные методы точны, хотя и медленны.

Калибровка по методу Tailor использует деление, а не вычитание, и это основывается на следующем эмпирическом наблюдении. Среднее значение нулевого распределения сильно коррелирует со стандартным отклонением, как показано на Рисунке 11 для двух экспериментальных наборов данных из нашего теста. Следовательно, нулевое распределение, которое имеет большее среднее значение, также имеет хвост, который затухает медленнее по сравнению с распределениями, которые имеют меньшее среднее значение. Следовательно, определенная разница, скажем, l, между лучшим показателем s_1 и $Q100$ ($l = s_1 - Q100$) может быть существенной для нулевых распределений с малым значением среднего, но может быть незначительной для тех, у кого большее среднее. Отношение s_1 к $Q100$ неявно учитывает

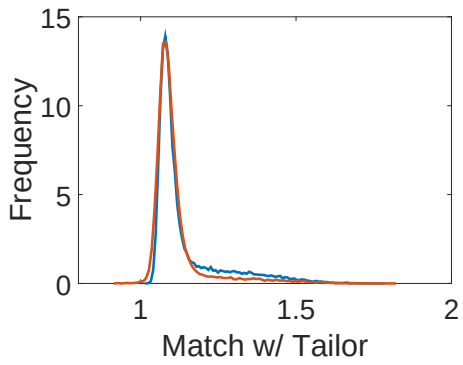

Рис. 12: Распределения PSM с нибольшими рейтингами, полученные при калибровке результатов сопоставления с помощью метода Tailor по данным Yeast для двух- (синих) и трех- (красных) зарядных родительских ионов.

размах нулевого распределения; то есть, чем шире нулевое распределение, чем выше его среднее значение, тем выше оценка $Q100$; следовательно, оценка s_1 откалибрована с большим фактором $Q100$. Таким образом, метод Tailor включает оценку разброса (std) нулевого распределения. Это и есть ключевое отличие от метрики XCorr.

Иллюстрация того, как калибровка Tailor уменьшает разницу между распределениями рейтингов двух- и трехзарядных спектров, показана на Рисунке 12.

4.4.1 Главные результаты BoltzMatch в области аннотации спектров

BoltzMatch был обучен с использованием трех наборов данных, полученных из предыдущих публикаций, которые содержали данные высокого разрешения MS2 (HumVar, iPRG, Malaria). Они содержали в общей сложности 41,792 спектра и были дискретизированы с шириной канала 0.05 Da. Мы подчеркиваем, что BoltzMatch был обучен и оценен по этим наборам данных отдельно, а результаты поиска BoltzMatch были откалиброваны по методу Tailor. Затем мы сравнили BoltzMatch с несколькими популярными поисковыми системами и визуализировали количество принятых PSM как функцию от q-значения на Рисунке 6. Результаты показывают, что BoltzMatch смог аннотировать 12,019 наблюдаемых спектров при 0.1% FDR, что на 49.05% больше аннотаций по сравнению со стандартной функцией оценки XCorr, в случае когда обе оценки были откалиброваны по методу Тейлора. И наоборот, XCorr аннотировал 12,019 спектров, содержащих в 5.3 раза больше ложных PSM, чем BoltzMatch. Метод Res-Ev с калибровкой XPV - это современная схема скоринга, разработанная специально для данных MS2 с высоким разрешением, которая была превзойдена BoltzMatch примерно на 17.78% аннотаций при FDR 0.1%; Res-Ev, напротив, дал 12,019 PSM с ошибками в 4.4 раза больше,

чем BoltzMatch.

4.5 Интерпретация BoltzMatch

Чтобы выявить причины, по которым BoltzMatch превосходит XCorr, преобразованный спектр $s_{BM} = s^T W$, полученный с весами BoltzMatch, сравнивался с s_{XC}, полученным с применением кросс-корреляционного штрафа XCorr [12], используя наблюдаемый спектр s из набора данных Malaria (scan id = 7990). Оба спектра показаны на Рисунках 13A-B.

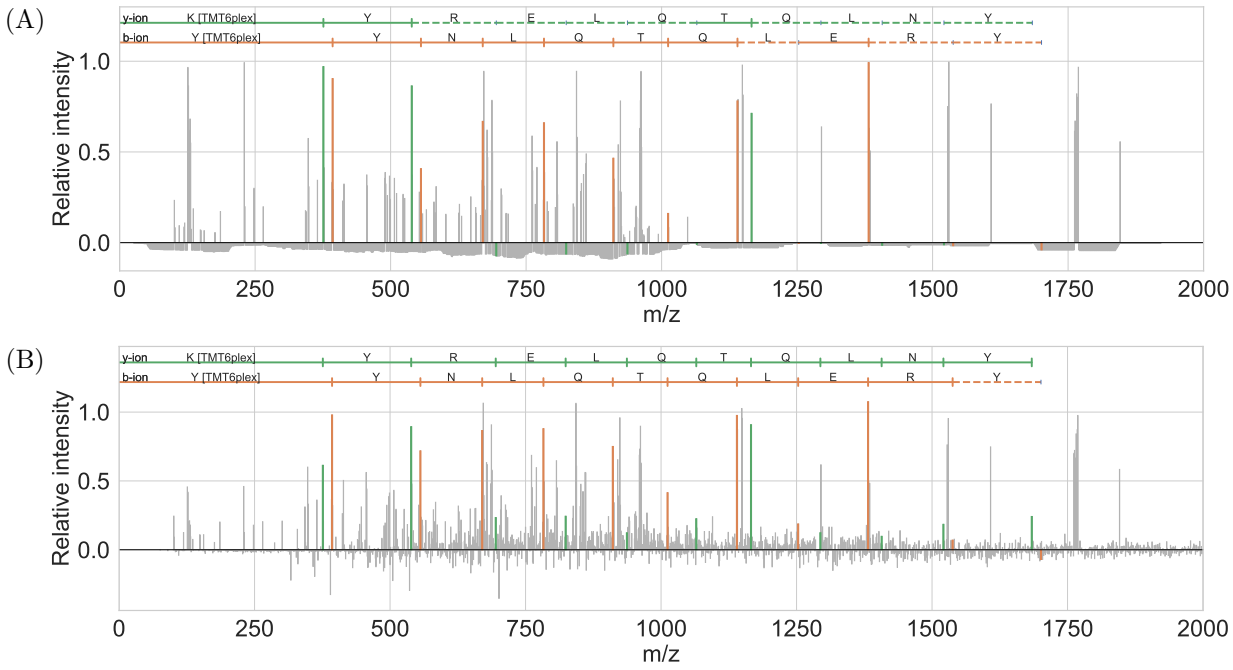

Рис. 13: Наблюдаемый спектр из набора данных Malaria (scan id = 7990), аннотированный с помощью теоретического пептида YYNLQTQLERY из базы данных. Пики с положительными значениями интенсивности, совпадающими с теоретическими $b-$ и $y-$ионами, отмечены зеленым и оранжевым цветами, соответственно. (A) Аннотация s_{XC}, полученная с помощью XCORR с q-значением 0.0049. (B) Аннотация s_{BM}, полученная с помощью BoltzMatch с q-значением 0.0019.

С одной стороны, этот рисунок показывает, что BoltzMatch нормализует интенсивность пиков в зависимости от того, могут ли они быть объяснено другими близлежащими пиками, тогда как XCorr уменьшает интенсивность пиков в зависимости от плотности соседних пиков независимо от семантики их контекста. Например, пик, соответствующий $b-$иону YY (красный пик в районе 600 m/z на Рисунке 13), был увеличен примерно на 55% с помощью BoltzMatch, но уменьшен на 10% с помощью XCorr. С другой стороны, BoltzMatch способен восстанавливать пики, соответствующие ненаблюдаемым, но ожидаемым фрагментарным ионам. Например, пик, соответствующий $y-$иону KYR (зеленый пик в районе 700 m/z на Рисунке 13), был восстановлен из соседних пиков в s_{BM} с положительным значением интенсивности, однако при трансформации XCorr он получает отрицательное значение интенсивности (то есть штраф) в s_{XC}.

4.6 Смещенные функции скоринга

Недавно в своем исследовании мы показали, что простая система скоринга, основанная на машинном обучении, может легко научиться отдавать предпочтение целевым пептидам, что, в свою очередь, приводит к смещенной оценке FDR [8], как это обсуждалось в Разделе 3.3.

Функции скоринга могут отдавать предпочтение целевым (или фиктивным) пептидам во время процедуры идентификации спектра, даже не учитывая класс пептида. Вопреки ожиданиям, распределение теоретических целевых и фиктивных спектров немного отличается в векторном пространстве спектров, и даже простая линейная модель может выучить эту информацию. Например, логистическая регрессия (LogReg) достигла 0.551 оценки AUC при классификации теоретических векторов целевого и фиктивного спектров, обученных и протестированных на полутриптических пептидах из белковых последовательностей Yeast, в которых фиктивные пептиды были получены путем reverse. Это означает, что функции скоринга, которые учитывают удельные веса, определенные для местоположения пика, могут вызывать смещение, независимо от того, настроены ли веса вручную или по определенному алгоритму машинного обучения. Ситуация с базовыми искусственными нейронными сетями (ANN) еще более удручающая (или удивительная). В том же наборе данных ANN достиг впечатляющей оценки AUC 0.902 для задачи классификации пептидов (см. синие и красные ROC-кривые на Рисунке 14). Это означает, что функции скоринга, учитывающие удельные веса пары пиков и местоположения пика, могут вызывать большие смещения. Возможно, методы глубокого обучения могут улучшить распознавание целевых и фиктивных пептидов между собой. Тем не менее, когда целевые (соответственно фиктивные) пептиды случайным образом разделяются на положительные и отрицательные группы, ANN достигает только 0.543 (соответственно 0.512) оценки AUC (см. Рисунки 3−5 в Приложении к статье "Bias in False Discovery Rate Estimation in Mass-Spectrometry-Based Peptide Identification"). По нашему мнению, это доказывает, что распределения целевого и фиктивного спектров действительно различны, и ANN не достигает высокого показателя AUC из-за "запоминания"данных.

На практике, например, метод скоринга DRIP [25] может отдавать предпочтение целевым пептидам. Основная проблема, по нашему мнению, является концептуальной. DRIP использует набор спектров, с размеченными как правильными целевыми пептидами, в качестве обучающего набора для обучения параметров динамической байесовской сети для моделирования правильного сопоставления экспериментальных спектров и пептидов. Тем не менее, во время обучения алгоритм также учит распределения целевых пептидов, что приводит к отдаванию им предпочтения при классификации. Мы показали это следующим экспериментом с использованием полностью триптических пептидов Yeast. Сначала мы сгенерировали 1000 пар пептидов и спектров in silico для обучения DRIP. Затем мы сгенерировали дополнительные 2000 синтетических спектров для поиска с помощью DRIP по несвязанному набору целевых и фиктивных пептидов. При таком подходе невозможно найти правильные PSM; таким образом, ожидается, что целевые и фиктивные пептиды с одинаковой вероятностью будут присваиваться масс-спектру. Функция скоринга XCorr дала на выходе случайные назначения предсказаний, в результате чего была получена оценка 0.496 AUC. Однако ROC-

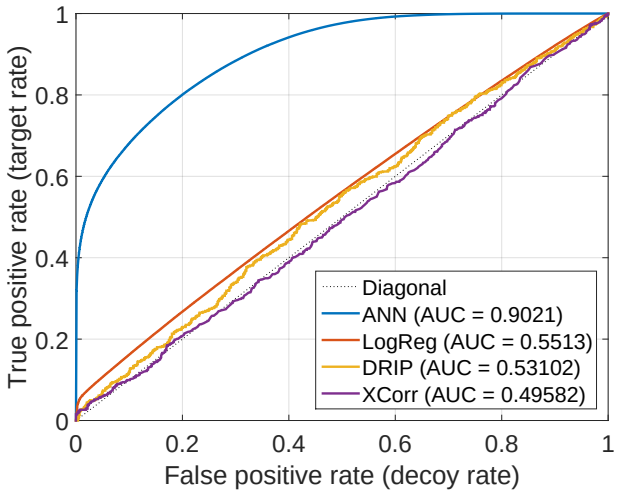

Рис. 14: Разделение целевых и фальшивых пептидов с помощью искусственной нейронной сети (ANN, синий), логистической регрессии (LogReg, красный), DRIP (оранжевый) и функции скоринга XCorr (фиолетовый), оцененной с помощью ROC-анализа. Диагональная линия (пунктирная линия) указывает на совершенно несмещенную функцию оценки, которая была бы идеальным вариантом.

анализ результатов поиска DRIP показал 0.531 AUC, что указывает на смещение в пользу целевых пептида (см. оранжевые и фиолетовые ROC-кривые на Рисунке 14).

4.6.1 Тестирование BoltzMatch на смещенность оценки

Здесь я утверждаю на теоретических и практических основаниях, что метод BoltzMatch не дает смещения.

Для теоретического обоснования рассмотрим экспериментальный спектр s. BoltzMatch будет давать смещение, если он назначит более высокие рейтинги целевым пептидам, чем фиктивным пептидам, то есть $\log p(s,t) \gtrapprox \log p(s,d)$ для независимо выбранных целевого t и фальшивого d пептидов, которые не связаны с s. Это будет означать, что распределение Гиббса, представленное ограниченной машиной Больцмана, имеет большую массу вокруг целевых пептидов, чем вокруг фиктивных пептидов. Однако целевые и фиктивные пептиды берутся из набора данных пептидов $t, d \in CP(S)$, и они в равной степени используются в $\sum_{h' \in CP(s)} p_w(s,h')\ s[j]\ h'[i]$ фазе (называемой негативной фазой). Это означает, что процедура обучения понижает ненормализованную вероятность сопоставления для t и d в равной степени, если они не связаны с s.

С практической точки зрения все 15,057 PSM с наибольшим рейтингом были взяты из набора данных HumVar, которые были получен с помощью оценки BoltzMatch, и были выбраны 5,000 PSM с наихудшим рейтингом (т.е. нижняя треть из списка ранжированных PSM) для дальнейший ROC-аналиа. Хвост ранжированного списка PSM должен содержать только неправильные аннотации спектра, которые с равной вероятностью могут быть сопоставлены либо с целевыми, либо с фиктивными пептидами в случае несмещенного метода оценки. Следовательно, распределение рейтингов целевых PSM (то есть PSM, в которых

спектры сопоставлены с целевыми пептидами) и распределение рейтингов фиктивных PSM должны быть неразличимы, что можно проверить с помощью ROC-анализа. Площадь под ROC-кривой (AUC), полученная с использованием описанных данных, отражена на Рисунке 15 и составляет 0.51, что соответствует p-значению 0.136, полученное с помощью двустороннего U-критерия Манна-Уитни. Это доказывает, что одно распределение рейтингов PSM стохастически не превышает другого при уровне значимости $\alpha = 0.1$. 2,000 PSM с наихудшими рейтингами в списке ранжированных PSM дают AUC 0.49927 с p-значением 0.95522.

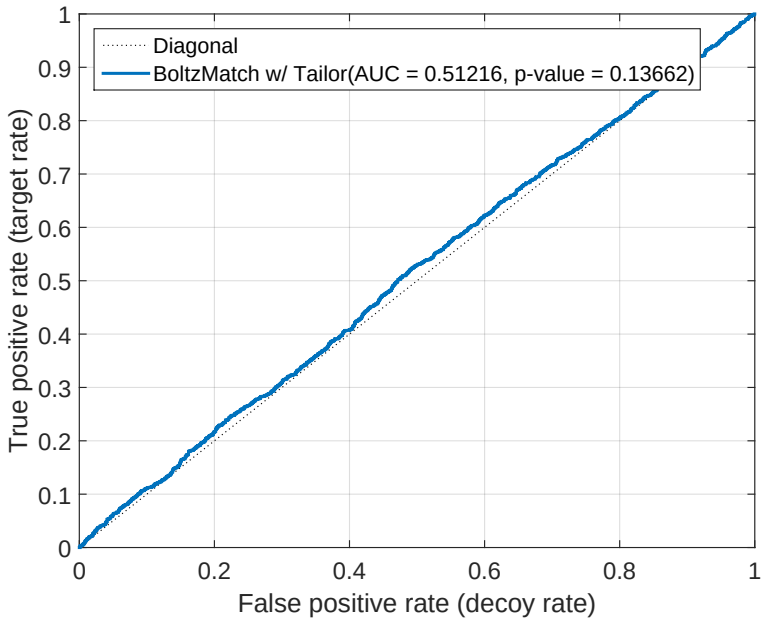

Рис. 15: ROC-анализ хвоста ранжированного списка PSM, полученного путем оценки набора данных HumVar с использованием обученного BoltzMatch. Диагональная линия (пунктирная линия) обозначает несмещенную функцию оценки и идентичные распределения целевых и фиктивных оценок PSM. Синяя линия показывает ROC-анализ распределения рейтингов целевых PSM по сравнению с распределением рейтингов фалшивых PSM. P-значение для ROC-анализа было получено с помощью двустороннего U-критерия Манна-Уитни.

5 Заключение

Многие, если не все, основанные на машинном обучении методы для идентификации масс-спектров используют обучение с частичным привлечением учителя. При таком подходе выполняется стандартный поиск по базе данных пептидов, чтобы получить некоторые аннотации спектров пептидами в качестве обучающих данных для метода машинного обучения, который, в свою очередь, может использоваться для получения, как мы надеемся, большего количества аннотаций спектров. Данный подход с использованием самообучения напоминает нам о "проблеме курицы и яйца": для аннотации большего количества спектров нужны аннотированные спектры. С одной стороны, это позволяет сразу создать набор обучающих данных для любого прибора и протокола эксперимента, используя быструю и простую функцию скоринга - и тогда остается вопрос, способен ли метод машинного обучения обобщать обучающие примеры для аннотирования большего количества спектров на любом уровне FDR.

Однако из-за недостаточного количества аннотаций человеком, в обучающих данных нет «железных примеров», показывающих сопоставления спектров и пептидов, которые действительно правильны, но неочевидны для объяснения и которые, вероятно, будут пропущены стандартными функциями скоринга. Таким образом, главный вопрос остается в том, можно ли обучать BoltzMatch без размеченных обучающих данных, т.е. можем ли мы обучать данный метод совершенно без учителя? Поскольку BoltzMatch берет свое начало из метода ограниченной машины Больтцмана, который может обучаться без привлечения учителя, в теории BoltzMatch можно обучать и без аннотированных данных спектров.

Я пытался обучать BoltzMatch без обучающих данных со следующими модификациями. При моделировании правильных соответствий $p(s, h)$, − где h были определены стандартными методами поиска по базе данных, как в уравнении 13 − h был заменен отфильтрованным экспериментальным спектром s'; то есть s' был получен из s путем удаления пиков с малой интенсивностью, то есть,возможно, шума. Подход привел к улучшению точности по сравнению с базовым уровнем, что означает, что методу удалось добиться большего обобщения по сравнению с базовым методом скоринга XCorr. К сожалению, этот подход не привел к дополнительным аннотациям по сравнению со случаем, когда использовалось частичное привлечение учителя. Однако, что касается будущих направлений исследований, было бы желательно исключить привлечение учителя в каком-либо виде, а также целевые и фальшивые пептиды из обучения, потому что это может снизить риск развития смещенных методов.

Список литературы

[1] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198, 2003.

[2] P. Alves, R.J. Arnold, M.V. Novotny, P. Radivojac, J.P. Reilly, and H. Tang. Advancement in protein inference from shotgun proteomics using peptide detectability. *Pacific Symposium On Biocomputing*, pages 409–420, 2007.

[3] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[4] K. Biemann, C. Cone, B.R. Webster, and G.P. Arsenault. Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *Journal of the American Chemical Society*, 88(23):5598—-5606, 1966.

[5] Wikimedia Commons. Schematic representation of a tandem mass spectrometry experiment., 2006. Own work of K. Murray.

[6] Jurgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: a peptide search engine integrated into the maxquant environment. *Journal of Proteome Research*, 10(4):1794–1805, 2011.

[7] Alan Crooks. Mass spectrometer, 2010. SlideShare.

[8] Yulia Danilova, Anastasia Voronkova, Pavel Sulimov, and Attila Kertész-Farkas. Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. *Journal of Proteome Research*, 18(5):2354–2358, 2019.

[9] Sven Degroeve and Lennart Martens. Ms2pip: a tool for ms/ms peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, 2013.

[10] Viktoria Dorfer, Peter Pichler, Thomas Stranzl, Johannes Stadlmann, Thomas Taus, Stephan Winkler, and Karl Mechtler. Ms amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research*, 13(8):3679–3684, 2014.

[11] Joshua E. Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 3(4):207–214, 2007.

[12] Jimmy K Eng, Bernd Fischer, Jonas Grossmann, and Michael J MacCoss. A fast sequest cross correlation algorithm. *Journal of Proteome Research*, 7(10):4598–4602, 2008.

[13] Jimmy K Eng, Michael R Hoopmann, Tahmina A Jahan, Jarrett D Egertson, William S Noble, and Michael J MacCoss. A deeper look into comet—implementation and features. *Journal of the American Society for Mass Spectrometry*, 26(11):1865–1874, 2015.

[14] Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. Comet: an open-source ms/ms sequence database search tool. *Proteomics*, 13(1):22–24, 2013.

[15] Jimmy K Eng, Ashley L McCormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[16] Dumitru Erhan, Pierre-Antoine Manzagol, Yoshua Bengio, Samy Bengio, and Pascal Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *AISTATS*, volume 5, pages 153–160, 2009.

[17] David Fenyö and Ronald C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, 75(4):768–774, 2003.

[18] Asja Fischer and Christian Igel. An introduction to restricted boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pages 14–36. Springer, 2012.

[19] Marjorie L. Fournier, Joshua M. Gilmore, Skylar A. Martin-Brown, and Michael P. Washburn. Multidimensional separations-based shotgun proteomics. *Chemical Reviews*, 107(8):3654–3686, 2007.

[20] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3(5):958–964, 2004.

[21] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster, and Mathias Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509–518, 2019.

[22] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[23] Viktor Granholm and Lukas Käll. Quality assessments of peptide–spectrum matches in shotgun proteomics. *Proteomics*, 11(6):1086–1093, 2011.

[24] Viktor Granholm, William Stafford Noble, and Lukas Käll. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *Journal of Proteome Research*, 10(5):2671–2678, 2011.

[25] John T. Halloran, Jeff A. Bilmes, and William S. Noble. Learning peptide-spectrum alignment models for tandem mass spectrometry. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 320—-329, Arlington, Virginia, USA, 2014. AUAI Press.

[26] Eric J Hartman, James D Keeler, and Jacek M Kowalski. Layered neural networks with gaussian hidden units as universal approximations. *Neural comp.*, 2(2):210–215, 1990.

[27] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.

[28] CS Ho, MHM Chan, RCK Cheung, LK Law, LCW Lit, KF Ng, MWM Suen, and Tai HL. Electrospray ionisation mass spectrometry: Principles and clinical applications. *The Clinical biochemist. Reviews*, 24(1):3–12, 2003.

[29] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[30] J Jeffry Howbert and William Stafford Noble. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Molecular & Cellular Proteomics*, 13(9):2467–2479, 2014.

[31] K. Ishikawa and Y. Niwa. Computer-aided peptide sequencing by fast atom bombardment mass spectrometry. *Biological Mass Spectrometry*, 13(7):373—-380, 1986.

[32] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923, 2007.

[33] H.R. Kaufman, United States. National Aeronautics, Space Administration, and Lewis Research Center. *Performance Correlation for Electron-bombardment Ion Sources*. NASA technical note. National Aeronautics and Space Administration, 1965.

[34] Uri Keich, Attila Kertész-Farkas, and William Stafford Noble. Improved false discovery rate estimation procedure for shotgun proteomics. *Journal of Proteome Research*, 14(8):3148–3161, 2015.

[35] Uri Keich and William Stafford Noble. On the importance of well-calibrated scores for identifying shotgun proteomics spectra. *Journal of Proteome Research*, 14(2):1147–1160, 2014.

[36] Attila Kertész-Farkas, Uri Keich, and William Stafford Noble. Tandem mass spectrum identification via cascaded search. *Journal of proteome research*, 14(8):3027–3038, 2015.

[37] Attila Kertész-Farkas, Beáta Reiz, Michael P Myers, and Sándor Pongor. Database searching in mass spectrometry based proteomics. *Current Bioinformatics*, 7(2):221–230, 2012.

[38] Sangtae Kim, Nitin Gupta, and Pavel A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of proteome research*, 7(8):3354–3363, 2008.

[39] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert JR Heck, and Pavel A Pevzner. The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840–2852, 2010.

[40] Sangtae Kim and Pavel A Pevzner. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[41] Aaron A Klammer, Christopher Y Park, and William Stafford Noble. Statistical calibration of the sequest xcorr function. *Journal of Proteome Research*, 8(4):2106–2113, 2009.

[42] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

[43] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *JMLR*, 10(Jan):1–40, 2009.

[44] Lev I. Levitsky, Mark V. Ivanov, Anna A. Lobas, and Mikhail V. Gorshkov. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *Journal of Proteome Research*, 16(2):393–397, 2017.

[45] Andy Lin, J Jeffry Howbert, and William Stafford Noble. Combining high-resolution and exact calibration to boost statistical power: A well-calibrated score function for high-resolution ms2 data. *Journal of Proteome Research*, 17(11):3644–3656, 2018.

[46] Bingwen Lu and Ting Chen. Algorithms for de novo peptide sequencing using tandem mass spectrometry. *Drug Discovery Today*, 2(2):85–90, 2004.

[47] Alexey I Nesvizhskii and Ruedi Aebersold. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug discovery today*, 9(4):173–181, 2004.

[48] Tran Ngoc Hieu, Zhang Xianglilan, Xin Lei, Shan Baozhen, and Li Ming. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31):8247–8252, 2017.

[49] William Stafford Noble and Michael J MacCoss. Computational and statistical analysis of protein mass spectrometry data. *PLoS computational biology*, 8(1):e1002296, 2012.

[50] David N Perkins, Darryl JC Pappin, David M Creasy, and John S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, 20(18):3551–3567, 1999.

[51] J.J. Pitt. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical Biochemist. Reviews*, 30(1):19–34, 2009.

[52] Donald Voet Pratt, G. Voet Judith, and W. Charlotte. *Fundamentals of biochemistry : life at the molecular level*. Hoboken, NJ: Wiley, 2006.

[53] Jesse Rodriguez, Nitin Gupta, Richard D. Smith, and Pavel A. Pevzner. Does trypsin cut before proline? *Journal of Proteome Research*, 7(1):300–305, 2008.

[54] B. Shin, H.J. Jung, S.W. Hyung, H. Kim, D. Lee, C. Lee, M.H. Yu, and S.W. Lee. Postexperiment monoisotopic mass filtering and refinement (pe-mmr) of tandem mass spectrometric data increases accuracy of peptide identification in lc/ms/ms. *Mol Cell Proteomics*, 7(6):1124–1134, 2008.

[55] H. B. Sieburg. Physiological studies in silico. *Studies in the Sciences of Complexity*, 12:321–342, 1990.

[56] Victor Spirin, Alexander Shpunt, Jan Seebacher, Marc Gentzel, Andrej Shevchenko, Steven Gygi, and Shamil Sunyaev. Assigning spectrum-specific p-values to protein identifications by mass spectrometry. *Bioinformatics*, 27(8):1128–1134, 2011.

[57] Pavel Sulimov and Attila Kertész-Farkas. Tailor: A nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics. *Journal of Proteome Research*, 19(4):1481–1490, 2020.

[58] Pavel Sulimov, Elena Sukmanova, Roman Chereshnev, and Attila Kertesz-Farkas. *Guided Layer-Wise Learning for Deep Models Using Side Information*, pages 50–61. Springer, 02 2020.

[59] Ilya Sutskever and Geoffrey E. Hinton. Deep, narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20:2629–2636, 2008.

[60] Craig D Wenger and Joshua J Coon. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of Proteome Research*, 12(3):1377–1386, 2013.

[61] M.R. Wilkins, R.D. Appel, K.L. Williams, and D.F. Hochstrasser. *Proteome Research Concepts, Technology and Application*. Springer, 2007.

[62] Dirk A. Wolters, Michael P. Washburn, and John R. Yates. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, 73(23):5683–5690, 2001.

[63] John R Yates, Jimmy K Eng, Ashley L McCormack, and David Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*, 67(8):1426–1436, 1995.

[64] Xie-Xuan Zhou, Wen-Feng Zeng, Hao Chi, Chunjie Luo, Chao Liu, Jianfeng Zhan, Si-Min He, and Zhifei Zhang. pdeep: Predicting ms/ms spectra of peptides with deep learning. *Analytical Chemistry*, 89(23):12690–12697, 2017.