NEWS
OF THENATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 2, Number 336 (2021), 56 – 65 **https://doi.org/10.32014/2021.2518-1726.21**

УДК 519.68.02

Е. С. Голенко, А. А. Исмаилова

Казахский агротехнический университет им. С. Сейфуллина, Нур-Султан, Казахстан.
E-mail: golenko.katerina@gmail.com, a.ismailova@mail.ru

СОВРЕМЕННЫЕ ВЫЧИСЛИТЕЛЬНЫЕ СТРАТЕГИИ ДЛЯ ВЫВОДА БЕЛКОВ В ПРОТЕОМИКЕ ДРОБОВИКА

Аннотация. Сегодня протеомика дробовика (Shotgun proteomics) является достаточно мощным подходом, дающим возможность охарактеризовать протеомы в биологических образцах. В отличие от стратегии протеомики «сверху вниз» протеомика дробовика характеризуется высокой эффективностью разделения и масс-спектральной чувствительностью. В то же время он предъявляет более высокие требования к вычислительным и статистическим методам, необходимым для идентификации пептидов, идентификации белков и количественного определения без меток. Основная цель протеомики дробовика – идентифицировать форму и количество каждого белка путем сочетания жидкостной хроматографии с тандемной масс-спектрометрией. Анализ и интерпретация экспериментальных данных являются заключительным и наиболее важным этапом в протеомике, они же порождают большое количество проблем, требующих сложных вычислительных решений. Одной из важнейших задач, безусловно, является идентификация белков, присутствующих в экспериментальном образце. Как правило, данную задачу подразделяют на две основные составляющие: этап присвоения экспериментальных тандемных масс-спектров пептидам, полученным из базы данных белков, и этап сопоставления пептидов с белками и количественной оценки достоверности идентифицированных белков. Также стоит принять во внимание, что оценка достоверности полученных данных может представлять собой отдельную, не менее важную и сложную задачу. В данной статье мы предлагаем рассматривать идентификацию белков не иначе как проблему статистического вывода, а также описываем ряд методов, которые могут быть использованы для её решения. Существующие подходы мы классифицируем на (1) методы, основанные на правилах, (2) комбинаторные методы оптимизации и (3) методы вероятностного вывода. Для представления методов используются целочисленное программирование и фреймворки байесовского вывода. Мы также обсуждаем основные проблемы идентификации белков и предлагаем возможные решения этих проблем.

Ключевые слова: протеомика дробовика, идентификация белков, масс-спектрометрия, LC/MS, вывод белков.

Введение. Основная цель протеомики на основе масс-спектрометрии – предоставить молекулярный снимок формы, уровня изобилия и функциональных аспектов каждого белка в биологический образец [1-3]. Среди протеомных стратегий восходящая или дроб-ная протеомика стала высокопроизводительной технологией, способной одновременно охарактеризовать сотни белков. В этом сценарии белки в образце сначала перевариваются в пептиды, обычно с использованием сайт-специфических протеолитических ферментов. Затем пептиды разделяют с помощью жидкостной хроматографии (ЖХ) и анализируют тандемной масс-спектрометрией (МС/МС), в результате чего получают набор спектров МС/МС [4].

В стандартном вычислительном конвейере МС/МС спектры масс-спектрометра ищутся по спектральным библиотекам [5] и/или in silico спектрам [6], соответствующим пептидам из базы данных белков, чтобы обеспечить совпадение пептидных спектров (Peptide-Spectrum Matches - PSM). Такой поиск в базе данных, в зависимости от парамет-ров поиска и платформы MS/MS, может привести к большому количеству PSM, которым присваиваются баллы, указывающие уровень достоверности правильной идентификации соответствующего пептида. Следующим шагом является составление списка идентифици-рованных белков из всех или подмножества PSM

и предоставление статистических уровней достоверности для каждого белка. Идентификация белков – это особый случай количественной оценки белков без меток, потому что в идеальном сценарии каждый белок с правильно выведенным ненулевым количеством будет считаться идентифицированным.

Получение списка идентифицированных белков из набора пептидных последователь-ностей с идентификационными баллами осложняется несколькими факторами: 1) обычно для каждого белка доступно лишь небольшое количество идентификаций пептидов, в большинстве своем ненадежные [7]. Это связано с тем, что только PSM с наивысшими баллами для каждого пептида обычно включаются в набор кандидатов для идентификации пептидов, и среди этих кандидатов только небольшая подгруппа считается достоверной идентификацией. Это приводит к трудностям в обеспечении надежной идентификации белков, например, если из белка идентифицируется только один пептид. 2) Пептиды, даже из одного и того же белка, вряд ли могут быть идентифицированы в протеомном эксперименте с одинаковой вероятностью [8]. Вероятность того, что пептид будет идентифицирован в стандартном протеомном эксперименте называется детектируемостью пептида [8]. 3) Многие пептидные последовательности, встречающиеся в типичном рабочем процессе протеомики, могут быть сопоставлены более чем с одним белком в базе данных. Их называют вырожденными или общими пептидами [9,10]. 4) Оценка частоты ложного обнаружения (False Discovery Rates - FDR) идентифицированных пептидов и белков также является весьма нетривиальной задачей. Некоторые подходы к оценке FDR на уровне пептидов включают создание ложных баз данных или неконтролируемую оценку условных распределений классов (распределения оценок PSM при правильной и ложной идентификации, соответственно). Однако большое количество PSM с низкой оценкой может создавать трудности в определении достоверности идентификации как пептидов, так и белков. Хотя методы оценки FDR на уровне пептидов хорошо описаны, вычисление FDR на уровне белков остается открытой проблемой [11, 12].

Процесс идентификации белков, присутствующих в биологическом образце, сегодня широко рассматривается и как проблема статистического вывода, и как проблема вывода белков [9, 10]. На сегодняшний день предложен ряд подходов для решения этой проблемы [9, 13-14]. Мы делим эти подходы на три большие группы:

1. Стратегии, основанные на правилах - методы, основанные на относительно неболь-шом наборе достоверно идентифицированных (уникальных) пептидов, которые впослед-ствии назначаются белкам.

2. Алгоритмы комбинаторной оптимизации - методы, которые полагаются на форму-лировки ограниченной оптимизации задачи вывода белков, приводящие, например, к минимальным спискам белков, которые покрывают некоторые или все достоверно иденти-фицированные пептиды.

3. Алгоритмы вероятностного вывода - методы, которые формулируют проблему вероятностно и назначают вероятности идентификации для каждого белка в базе данных.

В следующих разделах мы приводим обоснование разработки усовершенствованных алгоритмов вывода белков, а затем рассмотрим основные вычислительные стратегии. Все комбинаторные методы оптимизации представлены в рамках целочисленного программи-рования; с другой стороны, вероятностные алгоритмы резюмируются с использованием принципов байесовского вывода.

Основная часть. В большинстве экспериментов ЖХ-МС/МС, приводящих к потенциально большому количеству идентификаций белков, были сделаны выводы относительно влияния ошибочно идентифицированных белков на биомедицинскую науку. Это привело к созданию так называемого «правила двух пептидов» или правила двух совпадений, требующего наличия двух или более однозначно идентифицированных пептидов для определения достоверной идентификации белка [15]. Также рекомендовался принцип экономии в качестве объяснения достоверной идентификации пептидов и предлагалось, что «семейство белков» - белки со схожими последовательностями из-за вариантов одной аминокислоты, гомологов, вариантов сплайсинга или ошибок аннотации - следует рассматривать как одну группу, если белки имеют одни и те же идентифицирован-ные пептиды. В принципе, одного правильного уникального пептида должно быть достаточно для правильной идентификации белка. Однако даже для низкого FDR, связанного

с набором пептидов, многие отдельные пептиды в большом наборе данных идентифицированы неправильно. Более того, белки, идентифицированные с помощью одиночных попаданий пептидов, с большей вероятностью будут идентифицированы неправильно, чем белки с более высоким пептидным покрытием. Сообщалось, что FDR для белков с однократным попаданием могут быть более чем в 10 раз выше, чем FDR на уровне PSM [16], вероятно, из-за кластеризации правильных идентификаций пептидов с правильными белками и отсутствия поведения кластеризации для неправильных пептидов [16]. Позже правило двух пептидов было оспорено по нескольким причинам [17]. Во-первых, хотя включение белков с однократным попаданием без строгого контроля качества может поставить под угрозу специфичность, игнорирование таких белков, безусловно, снизит чувствительность [17]. Во-вторых, контроль достоверности (FDR) на уровне пептидов и последующее вычисление белков с использованием эвристических правил приводит к неопределенным FDR на уровне белков [16, 17]. С другой стороны, контроль FDR непосредственно на уровне белка может спасти некоторые из надежных белков с единичным попаданием. Gupta и Pevnzer продемонстрировали, что использование «правила одного пептида» приводит к увеличению количества идентификаций белков на 10-40% по сравнению с правилом двух пептидов при фиксированном уровне FDR [17]. Правило одного пептида просто использует пептид с наивысшей оценкой из белка в качестве балла для этого белка, а затем напрямую оценивает FDR на уровне белка (а не на уровне пептида) с использованием баз данных-приманок. Таким образом, любой белок, который имеет один или несколько пептидов с оценкой выше определенного порога, считается достоверным.

С помощью оценки FDR на уровне белка можно разработать более совершенные и более сложные правила для достижения еще более высокой чувствительности. Например, Weatherly и др. предложили установить отдельные пороги оценки для белков с разным количеством достоверных идентификаций пептидов. Для белков с более высоким охватом необходимо постепенно снижать пороговые значения. Для охвата 1 (то есть белков, пораженных отдельными пептидами) требовалось 44 балла по шкале MASCOT, а для охвата 6 баллов по шкале MASCOT всего 11 для того же FDR [18]. Несмотря на относительную простоту подходов, основанных на правилах, эффективность эвристических правил существенно ограничена из-за отсутствия строгой обработки и надлежащей комбинации баллов идентификации пептидов и предшествующих знаний.

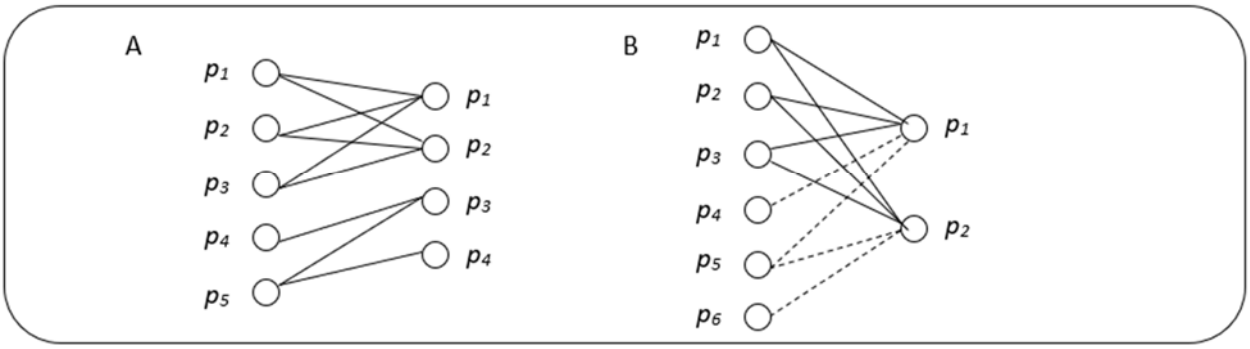
Комбинаторные алгоритмы оптимизации. Входные данные для этого класса алгорит-мов обычно состоят из набора достоверно идентифицированных пептидов $\mathcal{C} = \{p_j | t_j = 1\}$ и базы данных белков $\mathcal{P}$. Цель таких алгоритмов состоит в том, чтобы предоставить список белков, который оптимизирует определенные критерии. Как правило, все такие поста-новки приводят к NP-трудным задачам и обычно решаются с помощью приближенных алгоритмов.

Задача минимального покрытия множества (Minimum set cover (MSC) problem): имея набор достоверных идентификаций пептидов $\mathcal{U}$ и базу данных белков $\mathcal{P}$, необходимо найти наименьший список белков $\mathcal{L} \subseteq \mathcal{P}$, чтобы каждый пептид из $\mathcal{L}$ был назначен хотя бы одному белку из $\mathcal{L}$. Эта формулировка логического вывода белков идентична классической задаче информатики о минимальном покрытии множества, где при заданном наборе элементов (пептидов) $\mathcal{U}$ и наборе подмножеств (белков) над $\mathcal{U}$, цель состоит в том, чтобы найти наименьшее (не обязательно уникальное) множество подмножеств, которые содержат все элементы в $\mathcal{U}$. Формулировку MSC удобно визуализировать с помощью двудольных графов (рисунок А). Используя представление графа, относительно легко увидеть, что оптимальное решение проблемы MSC также может быть предоставлено, если исходный граф разделен на связанные компоненты, а оптимальное решение MSC предоставляется отдельно для каждого компонента.

Хотя состав MSC основан на наборе достоверно идентифицированных пептидов, предполагается, что подмножество таких пептидов будет неправильной идентификацией. Этот факт дает мотивацию для подходов к частичному покрытию множества, где цель состоит в том, чтобы найти минимальный список белков, который охватывает не менее $100 \cdot c\%$ идентифицированных пептидов, где $0 < c \leq 1$ - параметр, задаваемый пользователем.

Задача частичного минимального частичного покрытия множества (Minimum partial set cover (MPSC) problem): имея набор достоверных идентификаций пептидов $\mathcal{U}$, базу данных белков $\mathcal{P}$ и параметр c $(0 < c \leq 1)$, необходимо найти список белков $\mathcal{L}$ минимального размера такой, чтобы

по крайней мере $100 \cdot c\%$ идентифицированных пептидов относились к белкам из $\mathcal{L}$. Таким образом, оптимальные решения не могут быть гарантированы в ситуациях с большим количеством идентифицированных пептидов (каждый пептид из группы добавляет ограничение в формулировку задачи). Был предложен ряд алгоритмов аппроксимации, начиная от жадных алгоритмов до целочисленного программирования, и несколько таких алгоритмов были протестированы при выводе белков [19]. Формулировки задач MSC и MPSC приводят к ситуациям, когда невозможно различить белки, идентифицированные исключительно вырожденными пептидами (например, белки P1 и P2 на рисунке B). Nesvizhskii и Aebersold выделили несколько таких классов белков, назвав их неотличимыми белками, белками подмножества или объединяемыми белками [9].



Формулировка задачи о минимальном покрытии множества

Алгоритмы вероятностного вывода. Вероятностные подходы к выводу белков обычно состоят из двух этапов. Во-первых, оценки PSM преобразуются в вероятности PSM с использованием таких алгоритмов, как PeptideProphet. После этого этапа предвари-тельной обработки выполняется вывод белков на основе предполагаемой вероятностной модели. С вероятностной точки зрения, вывод белков включает вычисление апосте-риорных вероятностей $P(y_i = 1|\mathcal{S})$ для каждого белка в множестве $\mathcal{P}$. К настоящему времени было предложено несколько классов вероятностных алгоритмов [10,20-22] с разными стратегиями и уровнями строгости в обращении к группам белков и различной производительностью во время выполнения. Мы подробно обсуждаем три основных вероятностных метода: ProteinProphet [10], MSBayesPro [20] и Fido [22], а также кратко упоминаем несколько других методов, основной целью является возможность выявить внутренние связи и принципиальные различия между методами.

ProteinProphet — это первый и наиболее широко используемый подход вероятностного вывода белков [10], по важности сопоставимый с первым инструментом автоматической идентификации пептидов - алгоритмом SEQUEST. ProteinProphet состоит из четырех основных этапов; вместе они преобразуют исходные вероятности PSM из PeptideProphet в вероятности идентификации пептидов, а затем объединяют вероятности идентификации пептидов для вывода белков.

Предварительная обработка. Для получения вероятностей идентификации белков в качестве входных данных необходимы вероятности идентификации пептидов. Здесь трудность состоит в том, чтобы получить вероятность идентификации одного пептида из нескольких спектров, сопоставленных с пептидом. Решение, используемое в ProteinProphet, состоит в том, чтобы просто взять максимальное значение среди вероят-ностей совпадения пептидного спектра для пептида j, то есть (1):

$$P(x_j = 1|\mathcal{S}_j) = max_{s \in \mathcal{S}_j} P(x_j = 1|\mathcal{S}) \quad (1)$$

где $\mathcal{S}_j$ - набор спектров, идентифицированных для пептида j. Если пептиду не соответ-ствует ни один спектр, т. е. если $\mathcal{S}_j = \emptyset$, тогда $P(x_j = 1|\mathcal{S}_j) = 0$.

Комбинирование вероятностей пептидов. Ключевой особенностью ProteinProphet является то, что вероятности белков рассчитываются, исходя из предположения, что идентификация пептидов является независимым доказательством присутствия белка i в образце, то есть (2):

$$P(y_i = 1|\mathcal{S}) = \prod_{j \in N(i)} (1 - P(x_j = 1|\mathcal{S}_j)) \quad (2)$$

где $N\ (i)$ — это набор пептидов, сопоставленных с белком i. Это предположение, однако, нелегко обосновать, поскольку идентификация пептидов не является статистически независимой. То есть, если один пептид из белка идентифицирован достоверно, вероятность того, что другой пептид из того же белка также будет идентифицирован, выше. Другая проблема с этим предположением заключается в том, что каждый вырожденный пептид учитывается по отношению ко всем белкам, которым он соответствует. Эти проблемы решаются с помощью следующих двух шагов.

Корректировка вероятности идентификации пептида. Чтобы устранить ограничение, обусловленное предположением о независимости, ProteinProphet заменяет $P(x_j = 1|\mathcal{S}_j)$ в приведенном выше уравнении на $P(x_j = 1|\mathcal{S})$. Разница между скорректированной вероят-ностью идентификации пептида $P(x_j = 1|\mathcal{S}_j)$ и исходной вероятностью идентификации пептида $P(x_j = 1|\mathcal{S})$ происходит из-за наличия других спектров (пептидов), сопостав-ленных с тем же белком, что и пептид j. Ожидается, что спектры изменят достоверность идентификации пептидов.

Корректировка для пептидной вырожденности. Чтобы обратиться к вырожденным пептидам, используется схема взвешивания для изменения вероятностей белков (3):

$$P(\gamma_i = 1|\mathcal{S}) = \prod_{j\in N(i)}(1 - \omega_{ij} \cdot P(x_j = 1|\mathcal{S}_j)) \quad (3)$$

где ω_{ij} – «пропорция» пептида j, назначенная белку i. Nesvizhskii и др. определили, что $\omega_{ij} = P(\gamma_i = 1|\mathcal{S}) \,/\, \sum_{i'\in N(j)} P(\gamma_{i'} = 1|\mathcal{S})$, где $N(j)$ – множество белков, содержащих пептид j. Этот этап корректировки соответствует принципу экономичности $\sum_{i\in N(j)} \omega_{ij} = 1$, то есть каждый пептид гарантированно получен только из одного белка.

Как первый вероятностный метод вывода для идентификации белков ProteinProphet оказался очень успешным и, как часть Trans-Proteomic Pipeline [23] остается наиболее широко используемым инструментом вывода белков. Несмотря на то, что вырожденные пептиды обрабатываются с помощью процедуры взвешивания, основанной на принципе экономии, ProteinProphet использует итерационный метод для получения этих весов и, в конечном итоге, дает разумные вероятности для белков.

Однако, поскольку ProteinProphet полагается на некоторые сильные предположения, например, взвешивания, основанного на принципе экономии, его результаты не всегда разумны со статистической точки зрения. Например, для набора белков с общими пептидами белок с уникальным пептидом, независимо от того, насколько мала вероятность идентификации, всегда доминирует над белком (белками) без уникальных пептидов [21].

MSBayesPro — это байесовская сеть, служащая генеративной моделью для данных. Структура высокого уровня сети проста (Белки – Пептиды – Спектры) и имитирует экспериментальный протокол в протеомике, где белки сначала перевариваются в пептиды, из которых генерируются спектры. Следовательно (4),

$$P(\gamma, x, \mathcal{S}) = P(\gamma)P(x|\gamma)P(\mathcal{S}|x) \propto P(\gamma)P(x|\gamma)P(x|\mathcal{S}) \quad (4)$$

где γ - вектор случайных индикаторных переменных для всех белков-кандидатов, x - вектор случайных индикаторных переменных, представляющих все пептиды из этих белков, а $\mathcal{S}$ пред-ставляет данные, то есть все спектры, полученные в эксперименте. Связи Пептиды – Спектры определяются доступными показателями PSM (или вероятностями). Однако связи Белки – Пептиды определяются последовательностями пептидов и белков-кандидатов.

Модель MSBayesPro имеет важное свойство, заключающееся в том, что идентифи-кация пептидов условно независима, учитывая присутствие родительских белков. Это не следует путать с предположением о независимости идентификации пептидов, исполь-зуемым в ProteinProphet. Фактически, предположение об условной независимости в MSBayesPro приведет к незначительно зависимой идентификации пептидов, если два пептида имеют общие родительские белки прямо или косвенно через другие узлы пептида/белка (то есть, если два пептида находятся в связанном компоненте графа). Возможность обнаружения пептидов – еще одна важная отличительная черта MSBayesPro. Обнаруживаемость требуется для построения таблиц условного распределения между слоями белка и пептида и последующего вычисления апостериорных вероятностей для белков.

Однако для правильного использования обнаруживаемости важно учитывать влияние количества белка. Li и др. [24] предложили формулу корректировки количества для преобразования стандартной обнаруживаемости пептидов (5) в эффективную обнаруживаемость (6):

$$d_{ij}^{0} = P(x_i = 1|\gamma_i) = 1, q_i = q^0 \quad (5)$$

$$d_{ij}(q) = P(x_i = 1|\gamma_i) = 1, q_i = q \quad (6)$$

где q_i – количество белка P_i, которое оценивается методом максимального правдоподобия или подходов согласования моментов.

MSBayesPro использует выборку Гиббса вместо точных вычислений, когда связанный компонент в байесовской сети большой. Важно отметить, что MSBayesPro также сообщает оценочные количества белка и маргинальные апостериорные вероятности для пептидов, которые обеспечивают более высокие баллы для измерения достоверности пептидов [20]. Таким образом, по своей сути MSBayesPro также является алгоритмом количественной оценки без меток.

Использование возможности обнаружения пептидов является одновременно и сильной, и слабой стороной MSBayesPro. Другой недостаток связан с вычислительной сложностью: для работы MSBayesPro с очень большими наборами данных необходимы эффективные алгоритмы аппроксимации.

Модель Fido [22,25] использует байесовскую сеть, но в первую очередь была разра-ботана для быстрого вывода. Главный вклад этого метода состоит в двух преобразованиях графа, применяемых к каждому связанному компоненту: схлопыванию белковых узлов, которые связаны с идентичными наборами пептидов, и сокращению спектральных узлов (с параметрами, заданными пользователем), что приводит к разделению связанных компо-нентов. Также Fido позволяет применять передовые алгоритмы вероятностного вывода, например, алгоритм дерева соединений, который значительно улучшает вывод белков на больших графах.

Есть два основных различия в байесовских сетевых моделях, используемых Fido и MSBayesPro. Во-первых, неидентифицированные пептиды игнорируются в Fido, и параметр, не зависящий от последовательности, используется в качестве замены обнару-живаемости пептидов. Следовательно, результирующая байесовская сеть проще и быстрее. Во-вторых, в модель вводится еще один параметр b, который представляет собой априорную вероятность того, что пептид будет идентифицирован из искусственного «шумового» узла.

Одним из ограничений модели Fido является то, что она требует ложной (рандомизи-рованной) базы данных для нахождения лучших значений параметров (α, β и γ - априорность наличия белков) путем комбинирования ROC-оптимизации с оценкой FDR.

Выводы. Наша главная цель состояла в том, чтобы представить задачи, анализ и возможные решения проблемы вывода белков. В заключение мы обсудим текущие вопросы оценки алгоритмов вывода белков, а затем рассмотрим идеальные подходы к выводу белков.

Оценка методов идентификации белков. Несмотря на развитие вычислительных методов идентификации белков, объективная оценка эффективности методов остается открытой проблемой. В настоящее время доступны две стратегии: использование стандартных образцов (смесей известных белков) и использование последовательностей белков-ловушек для оценки FDR на уровне белка. Оба подхода имеют свои ограничения. Преимущество использования стандартных образцов в том, что заранее известна правда; таким образом, меры точности, например, точность и полноту идентификации белков можно вычислить напрямую. Однако стандартные образцы часто страдают от примесных белков, и граница между истинной и ложной идентификацией белка размыта. Еще одним ограничением стандартных образцов является небольшое количество белков, что приводит к трудностям в оценке статистической значимости при сравнении методов.

Второй подход позволяет оценить частоту ложных обнаружений на уровне белков с помощью ложных баз данных (баз данных-приманок или баз данных-ловушек). Мы предлагаем подходить к использованию ложных баз данных для оценки алгоритмов идентификации белков с учетом двух недостатков. Во-первых, в отличие от подхода к базе данных-приманок для пептидов, база данных-приманка для белков не дает правильной оценки количества неверных идентификаций белков, когда правильные белки составляют значительную часть базы данных. В крайнем случае, когда все

белки в базе данных присутствуют в образце, все идентифицированные белки из прямой базы данных верны, несмотря на то, что многие пептиды неверно идентифицированы. С другой стороны, все идентифицированные белки из базы данных-приманок неверны. Таким образом, использование приманки напрямую приведет к ненулевому FDR, тогда как FDR = 0 является правильным ответом.

Эту проблему можно решить, скорректировав смещение из-за количества истинных белков в прямой базе данных. Пусть количество идентифицированных прямых и ложных белков будет n_F и n_D, а общее количество прямых (forward) и ложных (decoy) белков в базах данных будет N_F и N_D соответственно. FDR уровня белка в прямой базе данных будет FDR_P, а частота неверных идентификаций белка из прямой и ложной базы данных будет (7) и (8) соответственно:

$$\gamma_F = \frac{FDR_P \cdot n_F}{N_F - (1 - FDR_P \cdot n_F)} \tag{7}$$

$$\gamma_D = \frac{)_D}{N_D} \tag{8}$$

В отношении базы данных-приманок предполагается, что частота ложных опреде-лений белков идентична, следовательно, $\gamma_F = \gamma_D$. Решая это уравнение, находим (9)

$$\gamma_D = \frac{n_D \cdot (N_F - n_F)}{n_F \cdot (N_D - n_D)} \tag{9}$$

Важно отметить, что в этом уравнении есть поправочный коэффициент $(N_F - n_F)(N_D - n_D)$. Кроме того, как и ожидалось, при $N_F = n_F, FDR_P = 0$. Важно отметить, что для алгоритмов вероятностного вывода белков теоретические значения FDR белка могут быть вычислены на основе апостериорных вероятностей белка. Однако такие теоретические значения FDR являются точными только тогда, когда сообщенные апостериорные вероятности белков верны.

Вторая и более серьезная проблема для применения метода приманки связана с существованием семейств белков. Рандомизированная база данных не может служить хорошей приманкой для оценки методов на наборах данных, которые содержат множество вырожденных идентификаций пептидов. Причина в том, что такие пептиды обычно являются общими для прямых белков, но не для белков-ловушек. В результате рандомизи-рованная база данных белков не может предоставить указания, правильны ли идентифи-кации, сделанные среди гомологичных белков. По этой причине ожидается, что рандомизированная база данных-приманок будет недооценивать FDR для образцов эукариот, которые имеют большое количество общих пептидов. Проблема может быть решена с использованием хорошо построенной базы данных неслучайных последователь-ностей или использования тесно связанной базы данных протеома в качестве приманки.

Необходимость руководящих принципов для сравнения методов. Во-первых, надежная и объективная проверка результатов идентификации белков сама по себе является сложной задачей, поскольку оценка FDR все еще ненадежна. Во-вторых, из-за отсутствия согласованных руководящих принципов в литературе иногда встречаются несправедливые сравнения, которых можно избежать [26].

Чтобы решить эту проблему, мы предлагаем следующие принципы для сравнения алгоритмов вывода белков. Во-первых, по возможности следует использовать одинаковые или эквивалентные баллы идентификации пептидов в качестве входных данных для разных программ. Во-вторых, необходимо приложить усилия для предоставления входных данных, наиболее подходящих для каждого рассматриваемого алгоритма. В-третьих, следует использовать, по крайней мере, один стандартный набор данных смеси белков, и все известные белки в таких наборах данных должны быть включены в оценку методов вывода белков. Это позволит оценить алгоритмы логического вывода белков для белков, идентифицированных без каких-либо уникальных пептидов. Наконец, в идеальном сценарии большие наборы данных из сложных образцов неизвестных белков также должны использоваться для сравнения различных программ.

Окончательный подход к выводу белков. Несмотря на количество опубликованных работ, проблема вывода белков далека от решения. Мы считаем, что два аспекта имеют решающее значение для будущих подходов. Во-первых, модель должна быть вероят-ностной и с принципиальным подходом к вырожденным пептидам. Во-вторых, неиденти-фицированные пептиды должны использоваться с возможностью обнаружения пептидов, включенной в модель. Определение белка можно рассматривать как частный случай количественной оценки без метки белка. Фактически, идеальный алгоритм вывода должен автоматически быть алгоритмом количественной оценки, и наоборот. Мы считаем, что гораздо лучших результатов можно добиться, объединив задачи по анализу белков и количественной оценке в одну статистическую структуру.

Е. С. Голенко, А. А. Исмаилова

С. Сейфуллин атындағы Қазақ агротехникалық университеті, Нұр-Сұлтан, Қазақстан

МЫЛТЫҚТЫҢ ПРОТЕОМИКАСЫНДАҒЫ АҚУЫЗДАРДЫ ЖОЮ ҮШІН ЗАМАНАУИ ЕСЕПТІК СТРАТЕГИЯЛАРЫ

Аннотация. Бүгінгі таңда мылтықтың протеомикасы - биологиялық үлгілердегі протеомдарды сипат-тайтын күшті тәсіл. Протеомика жоғарыдан төменге қарай стратегиясынан айырмашылығы, мылтық протеомикасы жоғары бөлу тиімділігімен және массаның спектрлік сезімталдығымен ерекшеленеді. Сонымен бірге, ол пептидтерді идентификациялауға, ақуыздарды идентификациялауға және этикеткасыз сандық анықтауға қажетті есептеу және статистикалық әдістерге жоғары талаптар қояды. Мылтық протеомикасының негізгі мақсаты - сұйық хроматографияны тандемді масс-спектрометриямен біріктіру арқылы әр ақуыздың пішіні мен мөлшерін анықтау. Эксперименттік мәліметтерді талдау және интерпрета-циялау протеомиканың соңғы және маңызды кезеңі болып табылады, сонымен қатар олар күрделі есептеу шешімдерін қажет ететін көптеген мәселелерді тудырады. Маңызды міндеттердің бірі, әрине, эксперименттік үлгідегі ақуыздарды анықтау. Әдетте, бұл тапсырма екі негізгі компонентке бөлінеді: ақуыздар базасынан алынған пептидтерге эксперименттік тандемдік масс спектрлерін беру кезеңі және пептидтерді ақуыздармен салыстыру және анықталған ақуыздардың сенімділігін сандық бағалау. Алынған мәліметтердің сенімділігін бағалау жеке, кем емес маңызды және күрделі міндет бола алатындығын ескерген жөн. Бұл мақалада біз ақуызды идентификациялауды тек статистикалық қорытынды мәселесі ретінде қарастыруды ұсынамыз, сонымен қатар оны шешуге болатын бірқатар әдістерді сипаттаймыз. Біз қолданыстағы тәсілдерді (1) ережеге негізделген әдістерге, (2) комбинаторлық оңтайландыру әдістеріне және (3) ықтималдық қорытын-дылау әдістеріне жіктейміз. Әдістерді ұсыну үшін бүтін бағдарламалау және байессиялық қорытынды жүйелері қолданылады. Сондай-ақ біз ақуызды идентификациялаудың негізгі мәселелерін талқылаймыз және осы мәселелерді шешудің мүмкін болатын жолдарын ұсынамыз.

Түйін сөздер: мылтықтың протеомикасы, ақуыздарды идентификациялау, масс-спектрометрия, LC/MS, ақуыздарды бөліп алу.

Y. S. Golenko, A. A. Ismailova

S. Seifullin Kazakh Agrotechnical University, Nur-Sultan, Kazakhstan

MODERN COMPUTATIONAL STRATEGIES FOR PROTEIN INFERENCE IN SHOTGUN PROTEOMIC

Abstract. Today, shotgun proteomics is a powerful approach to characterize proteomes in biological samples. Unlike the top-down proteomics strategy, shotgun proteomics is characterized by high separation efficiency and mass spectral sensitivity. At the same time, it places higher demands on the computational and statistical methods required for peptide identification, protein identification, and label-free quantification. The main purpose of shotgun proteomics is to identify the shape and amount of each protein by combining liquid chromatography with tandem mass spectrometry. The analysis and interpretation of experimental data is the final and most important stage in proteomics; they also generate a large number of problems that require complex computational solutions. One of the most important tasks, of course, is the identification of proteins present in the experimental sample. As a rule, this task is divided into two main components: the stage of assigning experimental tandem mass spectra to peptides obtained from the protein database, and the stage of comparing peptides with proteins and quantitative assessment of

the reliability of the identified proteins. It is also worth considering that the assessment of the reliability of the data obtained can be a separate, no less important and complex task. In this article, we propose to consider protein identification only as a problem of statistical inference, and also describe a number of methods that can be used to solve it. We classify the existing approaches into (1) rule-based methods, (2) combinatorial optimization methods, and (3) probabilistic inference methods. Integer programming and Bayesian inference frameworks are used to represent methods. We also discuss the main problems of protein identification and suggest possible solutions to these problems.

Keywords: shotgun proteomics, protein identification, mass spectrometry, LC/MS, protein inference.

Information about the authors:

Golenko Y.S., Doctoral Student, S. Seifullin Kazakh Agrotechnical University, Nur-Sultan, Kazakhstan; golenko.katerina@gmail.com; https://orcid.org/0000-0002-4643-4571;

Ismailova A.A., Ph.D., Senior Lecturer, S. Seifullin Kazakh Agrotechnical University, Nur-Sultan, Kazakhstan; a.ismailova@mail.ru; https://orcid.org/0000-0002-8958-1846

REFERENCES

[1] Aebersold R. and Mann M. (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537, 347–355. https://doi.org/10.1038/nature19949

[2] Cravatt B.F., Simon G.M., Yates J.R. (2007) The biological impact of mass spectrometry-based proteomics. Nature, 450(7172):991-1000. https://doi.org/10.1038/nature06525

[3] Choudhary C., Mann M. (2010) Decoding signaling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol, 11(6):427-439. https://doi.org/10.1038/nrm2900

[4] Steen H., Mann M. (2004) The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol, 5(9):699-711. https://doi.org/10.1038/nrm1468

[5] Lam H., Deutsch E.W., Eddes J.S., Eng J.K., Stein S.E., Aebersold R. (2008) Building consensus spectral libraries for peptide identification in proteomics. Nat Methods, 5(10):873-875. https://doi.org/10.1038/nmeth.1254

[6] Klammer A.A., Reynolds S.M., Bilmes J.A., MacCoss M.J., Noble W.S. (2008) Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics 24(13):i348-356. https://doi.org/10.1093/bioinformatics/btn189

[7] Resing K.A., Meyer-Arendt K., Mendoza A.M., Aveline-Wolf L.D., Jonscher K.R., Pierce K.G., Old W.M., Cheung H.T., Russell S., Wattawa J.L., et al. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal Chem, 76(13):3556-3568. https://doi.org/10.1021/ac035229m

[8] Tang H., Arnold R.J., Alves P., Xun Z., Clemmer D.E., Novotny M.V., Reilly J.P., Radivojac P. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics, 22(14):e481-e488. https://doi.org/10.1093/bioinformatics/btl237

[9] Nesvizhskii A.I., Aebersold R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics, 4(10):1419-1440. https://doi.org/10.1074/mcp.R500012-MCP200

[10] Nesvizhskii A.I., Keller A., Kolker E., Aebersold R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem, 75(17):4646-4658. https://doi.org/10.1021/ac0341261

[11] Elias J.E., Gygi S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 4(3):207-214. https://doi.org/10.1038/nmeth1019

[12] Nesvizhskii A.I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics, 73(11):2092-2123. https://doi.org/10.1016/j.jprot.2010.08.009

[13] Huang T., Wang J., Yu W., He Z. (2012) Protein inference: a review. Brief Bioinform. https://doi.org/10.1093/bib/bbs004

[14] Serang O., Noble W.S. (2012) A review of statistical methods for protein identification using tandem mass spectrometry. Stat Interface, 5(1):3-20. https://dx.doi.org/10.4310/SII.2012.v5.n1.a2

[15] Carr S., Aebersold R., Baldwin M., Burlingame A., Clauser K., Nesvizhskii A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol Cell Proteomics, 3(6):531-533. https://doi.org/10.1074/mcp.T400006-MCP200

[16] Reiter L., Claassen M., Schrimpf S.P., Jovanovic M., Schmidt A., Buhmann J.M., Hengartner M.O., Aebersold R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics, 8(11):2405-2417. https://doi.org/10.1074/mcp.M900317-MCP200

[17] Gupta N., Pevzner P.A. (2009) False discovery rates of protein identifications: a strike against the two-peptide rule. J Proteome Res, 8(9):4173-4181. https://doi.org/10.1021/pr9004794

[18] Weatherly D.B., Atwood J.A., Minning T.A., Cavola C., Tarleton R.L., Orlando R. (2005) A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Mol Cell Proteomics, 4(6):762-772. https://doi.org/10.1074/mcp.M400215-MCP200

[19] He Z., Yang C., Yu W. (2011) A partial set covering model for protein mixture identification using mass spectrometry data. IEEE/ACM Trans Comput Biol Bioinform, 8(2):368-380. https://doi.org/10.1109/TCBB.2009.54.

[20] Li and Radivojac. (2012) Computational approaches to protein inference in shotgun proteomics. BMC Bioinformatics 13(Suppl 16):S4. doi:10.1186/1471-2105-13-S16-S4

[21] Serang O., MacCoss M.J., Noble W.S. (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J Proteome Res, 9(10):5346-5357. https://doi.org/10.1021/pr100594k

[22] Shteynberg D., Deutsch E.W., Lam H., Eng J.K., Sun Z., Tasman N., Mendoza L., Moritz R.L., Aebersold R., Nesvizhskii A.I. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics, 10(12): M111 007690. https://doi.org/10.1074/mcp.M111.007690

[23] Deutsch E.W., Mendoza L., Shteynberg D., Farrah T., Lam H., Tasman N., Sun Z., Nilsson E., Pratt B., Prazen B., et al. (2010) A guided tour of the Trans-Proteomic Pipeline. Proteomics, 10(6):1150-1159. https://doi.org/10.1002/pmic.200900375

[24] Li Y.F., Arnold R.J., Li Y., Radivojac P., Sheng Q., Tang H. (2008) A Bayesian approach to protein inference problem in shotgun proteomics. The 12th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2008: 2008; Singapore, 167-180. https://doi.org/10.1007/978-3-540-78839-3_15

[25] Serang O., Noble W.S. (2012) Faster mass spectrometry-based protein inference: junction trees are more efficient than sampling and marginalization by enumeration. IEEE/ACM Transactions on Computational Biology and Bioinformatics. https://doi.org /10.1109/TCBB.2012.26

[26] Fengchao, et al: Identification of modified peptides using localization-aware open search Nat Commun. 2020 Aug 13;11(1):4065. https://doi.org/10.1038/s41467-020-17921-y