Қазақстан Республикасы Тәуелсіздігінің 30 жылдығына арналған «Сейфуллин оқулары – 17: «Қазіргі аграрлық ғылым: цифрлық трансформация» атты халықаралық ғылыми – тәжірибелік конференцияға материалдар = Материалы международной научно – теоретической конференции «Сейфуллинские чтения – 17: «Современная аграрная наука: цифровая трансформация», посвященной 30 – летию Независимости Республики Казахстан.- 2021.- Т.1, Ч.4 - С.215-218

МЕТОД ВЕРОЯТНОСТНОГО КОМБИНИРОВАНИЯ РЕЗУЛЬТАТОВ НЕСКОЛЬКИХ МЕТОДОЛОГИЙ ПОИСКА МС/МС ДЛЯ УВЕЛИЧЕНИЯ ВЕРОЯТНОСТИ ИДЕНТИФИКАЦИИ БЕЛКОВ

Голенко Е. С.
Исмаилова А. А.

Спектры, полученные при стандартном исследовании белков ЖХ/МС/МС, как правило, не полностью идентифицируются программами поиска в базах данных. Даже небольшие изменения в алгоритмах поиска в базе данных для отнесения пептидов к спектрам МС/МС могут давать разные результаты идентификации. Для того чтобы объединить результаты разных поисковых систем может использоваться вероятностная структура, при использовании которой баллы для каждой поисковой машины сначала независимо конвертируются в вероятности пептидов. Для объединения полученных вероятностей существуют различные стратегии, такие как алгоритм обучения максимизации ожидания и байесковские правила. С использованием каждой дополнительной системы поиска можно отследить увеличение количества корректно идентифицированных белков.

Процедура вероятностного объединения результатов идентификации пептидов из нескольких программ поиска в базе данных состоит из следующих шагов. Сначала для каждой поисковой системы создается модель вероятности пептида. Эта модель может оценить вероятность того, что любой индивидуальный спектр может быть назначен пептиду на основе дискриминантной оценки этой поисковой системы. Во-вторых, оценка согласованности поисковой системы вычисляется путем измерения степени согласованности между различными поисковыми инструментами для одного и того же спектра. В-третьих, различие между распределениями оценки согласия при поиске среди правильных и неправильных назначений пептидов определяется из данных. В-четвертых, индивидуальные вероятности пептидов поисковой машины корректируются на основе их согласованности с другими поисковыми системами. Наконец, для каждого спектра выбирается максимум этих вероятностей модифицированного пептида среди всех поисковых систем

как точная мера истинной вероятности того, что соответствующее назначение пептида является правильным [1].

Вычисление вероятностей пептидов для каждой программы поиска в базе данных. Программы поиска в базе данных присваивают спектры МС/МС пептидным последовательностям в базах данных белков и оценивают каждое присвоение. Присвоенные оценки могут быть преобразованы в вероятности с использованием подхода смешанной модели, реализованного в вычислительном инструменте PeptideProphet. В этом подходе вероятность того, что какой-либо конкретный спектр соответствует пептидной последовательности, оценивается путем рассмотрения гистограммы идентификационных баллов по всему эксперименту. Гистограмма обычно бимодальна, при этом высокие баллы обычно являются правильными (+), а низкие баллы - неправильными (-) с определенной степенью перекрытия. PeptideProphet выводит модель смеси для оценки природы этого перекрытия, подгоняя два распределения к оцененным правильным и неправильным спектрам на гистограмме. Подгонка распределений уточняется с использованием итеративного метода, называемого максимизацией ожидания.

Теорема Байеса представляет собой мощную основу для проверки гипотез, основанных на данных. Теорема вычисляет вероятность (p) того, что гипотеза верна $(+)$ с учетом записанных данных (D), или $p(+|D)$, где $|$ означает «с учетом допущений» [2]. Формальная формулировка теоремы Байеса такова (1):

$$p(+|D)=\frac{p(D|+)p(+)}{p(D)} \quad (1)$$

которая представляет собой вероятность того, что данные будут видны, когда гипотеза верна, или $p(D|+)$, умноженная на вероятность того, что гипотеза вообще может быть верной, или $p(+)$, деленная на вероятность того, что данные могут быть увиденным вообще, или $p(D)$. Если гипотеза может быть либо правильной, либо неправильной (т.е. либо пептид идентифицирован правильно, либо нет), то $p(D)$ можно переписать как (2):

$$p(+|D)=\frac{p(D|+)p(+)}{p(D|+)p(+)+p(D|-)p(-)} \quad (2)$$

Используя подобранные распределения, закон Байеса может помочь оценить вероятность того, что спектр МС/МС соответствует пептидной последовательности, $p(+|D)$, с учетом балла идентификации поисковой системы (D). Вероятность $p(D|+)$ представляет собой вероятность того, что совпадение с оценкой D присвоено правильному распределению, тогда как $p(D|-)$ — это вероятность того, что совпадение будет присвоено

неправильному распределению. Априорные вероятности $p(+)$ и $p(-)$ — это вероятности того, что любой точке данных присвоено правильное или неправильное распределение, независимо от оценки [3].

Простое сочетание программ поиска в базе данных. Взаимосвязь между каждым спектром (i) и пептидами (j), присвоенными поисковыми системами (k), проиллюстрирована на рисунке 1.

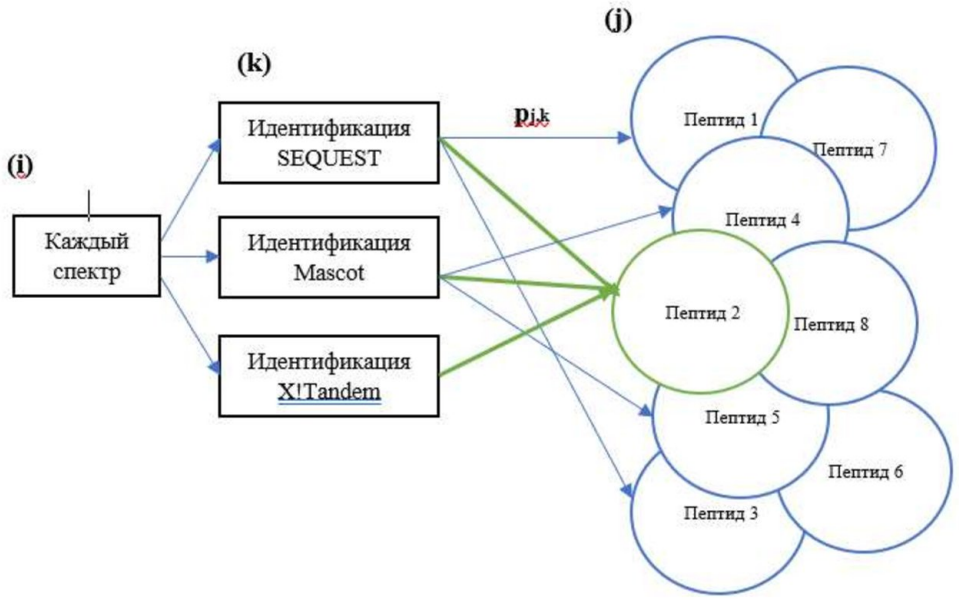

Рисунок 1 - Схема, показывающая, как спектр МС/МС может быть назначен пулу
пептидов несколькими поисковыми системами

Вероятность пептида для каждого спектра, присвоенного поисковой машиной k пептиду j, равна $p(+|D_{j,k})$. Комбинированная вероятность для всех поисковых систем равна $P(+|D_j)$. Если присвоение вероятности идентификации пептида разными поисковыми системами считается полностью независимым измерением одних и тех же данных, то вероятность того, что любая из поисковых систем правильно отнесет пептид j к нужному спектру i, будет равна (3):

$$P(+|D_j)=1-\prod_k(1-p(+|D_{j,k})) \quad (3)$$

Эта формула, вероятно, переоценивает вероятность, и консервативная оценка объединенной вероятности пептида для каждого пептида, $P(+|D_j)$, будет (4)

$$P(+|D_j)=max_k[p(+|D_{j,k})]$$

(4)

Комбинирование вероятностей пептидов с байесовской статистикой. Уравнение Байеса (2) может быть последовательно расширено для проверки гипотез, когда рассматриваются несколько типов данных, предполагая, что данные независимы. Формальное обозначение для этого (5):

$$p(+|D,A)=\frac{p(A|+)p(+|D)}{p(A|+)p(+|D)+p(A|-)p(-|D)} \quad (5)$$

(5) получается из (2) простой заменой $p(+)$ на $p(+|D)$ и $p(-)$ на $p(-|D)$. Этот метод можно использовать для улучшения оценки вероятности путем добавления дополнительной информации и будет полезен при объединении результатов из нескольких программ поиска в базе данных [4, 5].

Один из методов количественной оценки согласованности идентификации пептидов конкретного алгоритма поиска в базе данных с другими алгоритмами заключается в вычислении суммы вероятностей идентификации пептидов, которые другие программы сделали для того же пептида. Оценка соответствия для присвоения пептида j поисковой системой k спектру i определяется как (6):

$$A_{i,j,k}=\sum_{k'\neq k}\begin{Bmatrix} p(+|D_{i,j,k'}<0.05) & 00 \\ 0.05\leq p(+|D_{i,j,k'})<0.5 & 0.5 \\ 0.5\leq p(+|D_{i,j,k'}) & 1.0 \end{Bmatrix} \quad (6)$$

$A_{i,j,k}$ вычисляется для каждого назначения спектра по всем алгоритмам поиска в базе данных, за исключением k.

Идентификация белков с помощью нескольких поисковых систем. Комбинируя несколько инструментов поиска, на самом деле не удается идентифицировать больше белков как таковых [6]. Тем не менее, можно повысить уверенность во многих исходных идентификациях белков, которые в противном случае не прошли бы порог вероятности белка. Например, многие из добавленных белков были идентифицированы только одним пептидом, когда рассматривалась одна поисковая машина, но могли быть точно идентифицированы несколькими пептидами, когда использовалось несколько поисковых машин. В результате при фиксированной частоте ложных срабатываний общая чувствительность возрастает из-за повышенной уверенности. Одно интересное явление заключается в том, что распределение новых назначений пептидов согласовано для всех идентифицированных белков и не относится к какому-либо конкретному типу белка (данные не показаны). Это означает, что дополнительный охват спектра должен повысить

надежность количественных исследований подсчета спектра для всех белков [7].

Выводы. Использование нескольких поисковых машин по базам данных повышает надежность идентификации белков с более высоким охватом последовательностей. Кроме того, этот метод обеспечивает точные оценки вероятностей пептидов, а это означает, что результаты могут быть напрямую введены в алгоритмы вычисления вероятности белков, такие как ProteinProphet. Хотя за основу взяты только алгоритмы Mascot, SEQUEST и X! Tandem, метод может быть расширен для других инструментов идентификации. Наконец, расширение метода возможно и тем, чтобы включить результаты идентификации пептидов из других совершенно иных методов идентификации пептидов, таких как FingerPrint пептидной массы.

Список литературы

1. Sun, S. Improved validation of peptide MS/MS assignments using spectral intensity prediction / Sun, S., Meyer-Arendt, K., Eichelberger, B., Brown, R., Yen, C.Y., Old, W.M., Pierce, K., Cios, K.J., Ahn, N.G., and Resing, K.A. // Mol Cell Proteomics. – 2007. – № 6. – pp. 1-17.

2. Li Y.F. A Bayesian approach to protein inference problem in shotgun proteomics / Li Y.F. Arnold RJ, Li Y, Radivojac P, Sheng Q, Tang H. // The 12th Annual International Conference on Research in Computational Molecular Biology, RECOMB. – 2008. – pp. 167-180.

3. Wisniewski, J.R. A Proteomics Approach to the Protein Normalization Problem: Selection of Unvarying Proteins for MS-Based Proteomics and Western Blotting / Wisniewski, J.R., Mann, M. // J. Proteome Res. – 2016. – №15. – pp. 2321–2326.

4. Serang O, Noble WS. (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.

5. Fengchao, et al. Identification of modified peptides using localization-aware open search / Fengchao, et al. // Nat Commun. – 2020. – №11. – pp. 4065.

6. Rudnick, P.A. Large scale analysis of MASCOT results using a mass accuracy-based THreshold (MATH) effectively improves data interpretation / Rudnick, P.A., Wang, Y.J., Evans, E., Lee, C.S. & Balgley, B.M // J. Proteome. – 2005. – №4. – pp. 1353–1360.

7. Keller A. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search / Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, Ruedi Aebersold // Anal Chem. – 2002. – №74. – pp. 5383-92.