Московский физико-технический институт (государственный университет)
Факультет биологической и медицинской физики
Кафедра кафедра молекулярной и трансляционной медицины

Диссертация допущена к защите
зав. кафедрой

_____ Лазарев В.Н.

«_____» _____ 2017 г.

Выпускная квалификационная работа на соискание степени МАГИСТРА

Тема: **Количественный протеогеномный анализ туберкулеза и ещё чего-нибудь**

Направление: 010900 – Прикладные математика и физика
Магистерская программа: 010982 – Физико-химическая биология и биотехнология

Выполнил студент гр. 1114 _____ Смоляков А.В.

Научный руководитель,
к. б. н. _____ Шитиков Е.А.

Работа выполнена в ФГБУ ФНКЦ ФХМ ФМБА России
Москва – 2017

Оглавление

1. Список сокращений

GSSP - Genome Search Specific Peptides. Это пептиды, индетифициуремые при поиске против геномной базы, и не идентифицируемые при происке против протеомной.

2. Введение

3. Обзор литературы

3.1. Mycobacterium tuberculosis

3.2. Применение масс-спектрометрии в протеомике

3.3. Orbitrap

3.4. Протеогеномика

Для получения протеомных данных обычно используют метод «shotgun proteomics» - комбинация жидкостной хроматографии и тандемной масс-спектрометрии [1]. Одним из ключевых шагов в протеомике является идентификация пептидов на основе полученных MS/MS спектров. В отличее от геномных технологий, вроде ДНК или РНК секвенирования, где происходит непосредственное секвенирование исходной последовательности, в протеомике, как правило, пептиды идентифицируются за счет сопоставления экспериментальных MS/MS спектров и теоретических спектров всех пептидов, представленных в базе, против которой осуществляется поиск [2]. При таком поиске используются следующие исходные предположения: 1. все белок-кодирующие последовательности генома точно известны и аннотированы 2. все эти последовательности включены в базу, против которой осуществляется поиск . Весь последующий анализ, включая идентификацию, количественный анализ и прочий статистический анализ, основывается на этих предположениях [3].

Проблема такого подхода заключается в том, что не все пептиды представлены в текущей поисковой базе или какой-либо другой. Пептиды могут содержать мутации, находиться в новых генах, перед неверно аннотированным стартом или в альтернативных сплайсформах. Один из способов идентификации пептидов с мутациями заключается в масс-тег подходе. При этом подходе происходит идентификация коротких участков пептида, после чего осуществляется поиск в более широком диапазоне масс прекурсеров [4].

Более общим подходом является протеогеномика. Термин впервые был использован в 2004 и изначально использовался в исследовании, где протеомные данные использовались для улучшения качества аннотации [5]. С тех пор этот термин используется в более общем смысле. В протеогеномном подходе, пептиды идентифицируются за счет идентификации MS/MS спектров против специальной базы, включающей

в себя последовательности новых, предсказанных белков и различные варианты последовательности белка. Такие базы получаются за счет использования геномной и транскриптомной информации. Таким образом, протеогеномика позволяет не только подтвердить текущую аннотацию, но так же уточнить её [6].

Типы пептидов, идентифицируемых при протеогеномном исследовании

Пептиды, идентифицируемые при протеогеномном поиске, соответствуют различным участкам генома. Такие пептиды можно разделить на межгенные (находятся между аннотированными генами) и внутрегенные (находятся полностью или частично в областях, где содержится аннотированный ген). Внутрегенные можно разделить на 1. находящиеся в белок-кодирующих генах 2. находящиеся в длинных некодирующих РНК 3. находящиеся в псевдогенах [7]. Большинство пептидов, идентифицируемых в протеогеномике, уже известны и относятся к аннотированным генам. В эукариотах (в которых присутствует интроно-экзоная структура) большинство пептидов относятся к экзомам, и, как правило, меньше 20% относятся к экзон-экзон участкам. Новые пептиды, не идентифицируемые против какой-либо базы, могут находится в неаанотированных участках генома, быть результатами одно-аминокислотной замены (SAP), находится в нетранслируемых регионах (3' или 5' UTR) или интронах, является результатам альтернативного сплайсинга [6].

Подходы к созданию поисковых баз

Идентификация пептидов против против кастомных баз является ключевым шагом в протеогеномике. Обычно база состоит из известных аннотированных последовательностей и предсказанных последовательностей. При протеогеномном поиске следует внимательно относиться к размеру базы: увеличение размера влечет за собой увлечение времени поиска и FDR. Оптимальный выбор зависит от того, что требуется в эксперименте: точность или чувствительность [6].

Транслирование в шести рамках генома - такая база может получена в результате транслирования в шести рамках генома [8]. Недостатком такого подхода является гигантский размер итоговой базы (в основном состоящей из последовательностей несуществующих белков) и невозможности поиска экзом-экзом пептидов, в случае эукариот. Например транслированный таким образом геном человека приво-

дит к базе в 3.2 гигабазы белковых последовательностей, что в 70 раз больше, чем референс в 45 мегабаз [9]. Для уменьшения размера базы могут применены различные вычислительные методы: выбор последовательностей, имеющих гомологии с уже известными белками; использование методов предсказания кодирующего потенциала; исключение слишком коротких последовательностей (например, меньших, чем 30 аминокислот) [10].

Ab initio предсказание генов

Expressed sequence tag (EST) data

Аннотированные РНК-транскрипты Белковые последовательности могут быть получены в результате шестирамочного транслирования аннотированных РНК-транскриптов, например Enselbl или RefSeq. Это позволяет идентифицировать альтернативные сайты инициации трансляции. База GENCODE содержит 84408 мРНК аннотированных белков. В результате транслирования такой базы получается белковая база в 200 мегабазы, что всего в 4.5 раза больше референса [9]. Так же такие базы могут содержать последовательность, аннотированные как псевдогены или длинные некодирующие РНК [11].

RNA-seq данные

Различные вариации последовательностей Белковые последовательности в рефернсной базе могут быть дополнены последовательностями, являющими вариациями референсных последовательностей (как правило, это одно аминокислотные полиморфизмы, делеции и инсерции). Для каждой вариации, берется большая область вокруг вариации и добавляется в базу, как независимая последовательность. Одно аминокислотные замены можно скачать из различных баз данных: NCBI dbSNP, Online Mendelian Inheritance, Protein Mutant Database [12].

Прочие специализированные базы

3.5. Влияние размера базы

Возможность идентифицировать спектры пептидов, полученные в результате MS/MS экспериментов, используя поисковые базы данных, зависит от многих факторов. Во-первых, пептид должен присутствовать в поисковой базе. Однако, чем больше сравнивается теоритических спектров с экспериментальным, тем больше вероятность того, что лучший результат будет у неверного теоритического спектра, и

тем труднее различать верные и неверные идентификации [2]. В результате, поиск против большой базы может дать несколько новых идентификаций белков или пептидов, но при этом общее количество идентификаций будет меньше, в сравнении с поиском против референсного сиквенса [10, 13]. Так же увелечение базы приводит к увелечению машинного времени, необходимого для этого поиска. Таким образом, одним из ключевых моментов в протеогеномике является поиск баланса между размером базы и её содержимым.

3.6. Способы увелечение чувствительности идентификации пептидов

Подходы, используемые в протеомике для увелечения числа идентификаций, включают в себя идентификацию с помощью нескольких поисковых машин одно и того же набора данных [14]. После поиска происходит изменение скорингов иднетификации пептидов за счет комбинации нескольких источников информации, используя подходы машинного обучения [2]. Одной из дополнительных стратегий для сокращения простанства поиска является фракционирование смеси пептидов, проводимое до LS-MS/MS анализа. Фракционирование может проходить за счет опредленных физико-химических свойств пептида или свойств последовательности пептида. Примером фракционирования может служить изоэлектрическая фокусировка. Спектры, полученные от фракции с опредленной изоэлектрической точкной, можно искать против пептидов с примерно такой же теоритически предсказанной изоэлетктрической точкой [15].

Чувствительность идентификации так же можно повысчить за счет многоступенчатого анализа данных. На первом этапе можно проводить поиск против “эталонной” базы, наиболее точно описывающей исслдеуемый организм и позволяющей идентифицировать большинство спектров. На втором этапе происходит поиск против расширенной базы для идентификации дополнительных спектров [16, 17].

Поиск новых генов и корректировка рамок

Причины приводящие к неточности аннотации

4. Материалы и методы

Общая схема эксперимента приведелена на рисунке 1.

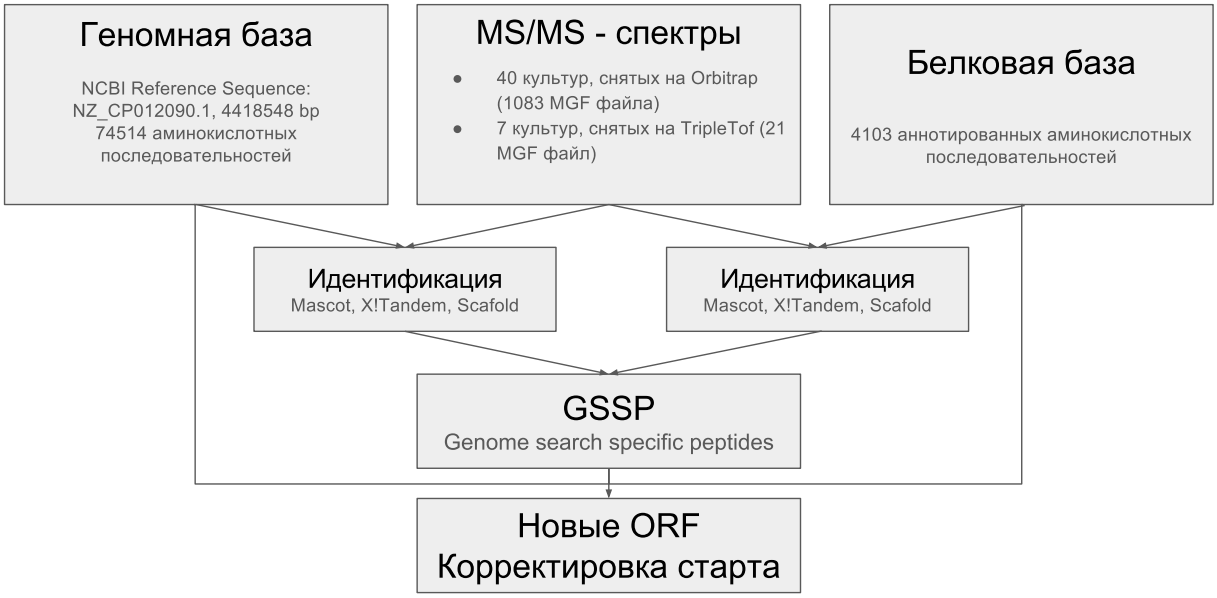

Рис. 1. Схема анализа масс-спектрометрических данных. Можно выделить следующие ключевые шаги: *a)* идентификация против белковой базы *b)* идентификация против геномной базы *c)* нахождение GSSP *d)* интерпретация GSSP.

4.1. Получение бактерий

Бактерии были выделены из мокроты, больных туберкулёзом пациентов.

4.2. Проведение масс-спектрометрического эксперимента

Для протеогеномного анализа использовались масс-спектры белковых клеточных лизатов для *M. tuberculosis*, полученных с прибора AB SCIEX TripleTOF 5600 в лаборатории протеомного анализа НИИ ФХМ и ThermoFisher Q Exactive Hybrid Quadrupole-Orbitrap в ИБМХ.

4.3. Контроль качества

Для всех масс-спектров был проведен контроль качества масс-спектров с использованием программного решения реализованного в лаборатории биоинформатики НИИ ФХМ. В ходе контроля качества были проверены следующие факторы:

качество трипсинолиза, распределение зарядов родительских ионов, ошибка измерения m/z для родительских и дочерних ионов, распределение идентифицированных пептидов по времени удерживания пептидов в хроматографической колонке.

4.4. Создание поисковых баз

В работе использовалось 2 типа баз: белковая и геномная. Белковая база - аннотированные последовательности, для данного штамма. Геноманя - база, полученная в результате транслирования генома в шести рамках. Белковые базы для *M.tuberculosis W-148* и *M.tuberculosis H37Rv* были составлены из аннотированных белков штаммов (*W-148*: NCBI Reference Sequence: NZ_CP012090.1, версия от 11 марта 2017 года, 4103 аминокислотных последовательности, 137 псевдогена; *H37Rv*: NCBI Reference Sequence: NC_000962.3, версия от 2 августа 2016 года, 3932 аминокислотных последовательности). Геномные базы были получены в результате 6 рамочного транслирования от стоп- до стоп-каднона геномов штаммов *M.tuberculosis W-148* и *M.tuberculosis H37Rv*, используя программу Artemis версия 16.0.0 [18]. При трансляции использовалась 11 трансляционная таблица NCBI. Минимальная длинна рамки была установлена в 100 нуклеиновых кислот. К каждой базе были добавлены последовательности 26 контаминантных белков (кератины, альбумины, трипин) и decoy-последовательности, полученные в результате прочтения аминокислотных последовательностей с конца, за исключением стартового метеонина.

4.5. Идентификация пептидов и белков

Данные полученные в результате LC-MS/MS эксперимента (Raw формат) были сконвертированы в пик-лист (MGF формат), используя ProteoWizard msconvert [19]. Идентификация проходила против двух белковых и двух геномных баз с использованием Mascot Search Engine version 2.5.1 [20] и X!Tandem version 3.4.3 [21]. Результаты идентификации двух программ были объеденины в Scaffold version 4.2.1.

Параметры поиска Mascot были следующими: триптические пептиды, не более одного пропущенного сайта трипсинолиза, ошибка массы прекурсера 20 ppm, ошибка массы фрагментов 0.04 Да, заряды прекурсера 2+, 3+, 4+. Oxidation(M) была установлена как возможнная модификация пептидов, Carbamidomethylation(C) как фиксированная.

Параметры X!Tandem были следующими: триптические пептиды, не более одного пропущенного сайта трипсинолиза, ошибка массы прекурсера 20 ppm, ошибка массы фрагментов 50 ppm, проверка не моноизотопных масс, Carbamidomethylation(C) - фиксированная модификация, Oxidation(M) возможная модификация.

Результаты работы поисковых машин были объединины в Scaffold с параметрами: 1:1 forward/decoy ratio, LFDR scoring, стандартные белковые группы, не проводить GO-аннотацию. Белковый и пептидный FDR был установлен на уровне 1%, 1 и более пептидов на белок. Результаты были экспортированны в виде листа идентифицированных пептидов.

4.6. Протеогеномика *W-148*

Координаты аннотированных генов были пересечены с учетом стренда и фрейма с координатами ORF, полученными в результате шестирамочного транслирования. Для поиска GSSP из результатов поиска против геномной базы *W-148* были исключены пептиды, идентифицированные против белковой базы *W-148*. Так же были исключены пептиды идентифицируемые против геномной базы и предсталвенные в аннотации, и пептиды, идентифицируемые только в одной культуре. Для дальнейшего анализа были выбраны ORF, в которых произошло одно из следующих событий: 1. идентифицированно два и более уникальных GSSP 2. идентифицированн GSSP и присутствует аннотированный ген 3. идентифицированн GSSP и есть пересечение по координатам с псевдогеном в пределах стренда .

Валидация результатов идентификации

Для каждого GSSP была проведена проверка времени выхода и ошибки масс при идентификации; потенциальной остаточной контаминации на приборе; ошибки интерпретации модифицированной акминокислоты, как другой немодифицированной.

Проверка ошибки идентификации масс проходила на уровне PSM. В каждом ране, для каждого PSM, соответсвующему GSSP, находилась среднее значение и стандартное отклонение ошибки идентификации всех PSM в диапазоне +/- 5 минут от времени выхода данного PSM. Из дальнейшенго анализа исключались все PSM, ошибка идентификации масс которых отличалась более чем на 3 стандартных откло-

ения от средей ошибки в установленном временом интервале. В каждом ране были отфильтрованны 5% самых ранних и поздних по времи выхода PSM.

Был проведен поиск точного и с учетом одной возможной замены вхождения GSSP в другие белки, предсталвенные в базе NCBInr.

Идентификация новых белков

Рассматривались ORF, в которых было идентифицированно два и более уникальных GSSP-пептида и в которых не содержится аннотиванный ген. Для проверки потенциала кодирующей способности рамки, был проведён blastp против базы nr.

Уточнение N-концов

Рассматривались ORF, в которых было идентифицированно два и более уникальных GSSP-пептида и в которых содержится аннотиванный ген. Новые рамки сравнивались с аналогичными генами в *H37Rv*.

4.7. Визуалиция данных

Для визуализации данных использовался Gbrowse. Были выделены следующие глифы: 1. аннотированные гены 2. идентифицированные пептиды 3. псевдогены 4. ORF с новыми генами 5. ORF с пептидами, идентифицируемые перед аннотированным стартом 6. GSSP-пептиды. Результаты идентификации были обработаны и экспортированны в gff3 формате.

5. Результаты и обсуждение

5.1. Протеогеномика *W-148*

Идентификация

Новые гены и их валидация

Уточнение N-концов

5.2. Сравнение идентификаций против *W-148* и *H37Rv*

Новые гены и их валидация

Уточнение N-концов

Валидация SAP

6. Выводы

Список литературы

1. Bantscheff M., Lemeer S., Savitski M. M., Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present // Analytical and bioanalytical chemistry. 2012. Vol. 404, no. 4. P. 939–965.
2. Nesvizhskii A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics // Journal of proteomics. 2010. Vol. 73, no. 11. P. 2092–2123.
3. Nesvizhskii A. I., Aebersold R. Interpretation of shotgun proteomic data the protein inference problem // Molecular & Cellular Proteomics. 2005. Vol. 4, no. 10. P. 1419–1440.
4. Dasari S., Chambers M. C., Slebos R. J. et al. TagRecon: high-throughput mutation identification through sequence tagging // Journal of proteome research. 2010. Vol. 9, no. 4. P. 1716.
5. Jaffe J. D., Berg H. C., Church G. M. Proteogenomic mapping as a complementary method to perform genome annotation // Proteomics. 2004. Vol. 4, no. 1. P. 59–77.
6. Nesvizhskii A. I. Proteogenomics: concepts, applications and computational strategies // Nature methods. 2014. Vol. 11, no. 11. P. 1114–1125.
7. Harrow J., Frankish A., Gonzalez J. M. et al. GENCODE: the reference human genome annotation for The ENCODE Project // Genome research. 2012. Vol. 22, no. 9. P. 1760–1774.
8. Baerenfaller K., Grossmann J., Grobei M. A. et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics // Science. 2008. Vol. 320, no. 5878. P. 938–941.
9. Khatun J., Yu Y., Wrobel J. A. et al. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions // BMC genomics. 2013. Vol. 14, no. 1. P. 141.
10. Blakeley P., Overton I. M., Hubbard S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies // Journal of proteome research. 2012. Vol. 11, no. 11. P. 5221–5234.
11. Derrien T., Johnson R., Bussotti G. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression //

Genome research. 2012. Vol. 22, no. 9. P. 1775–1789.

12. Li J., Su Z., Ma Z.-Q. et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics // Molecular & Cellular Proteomics. 2011. Vol. 10, no. 5. P. M110–006536.
13. Krug K., Carpy A., Behrends G. et al. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments // Molecular & Cellular Proteomics. 2013. Vol. 12, no. 11. P. 3420–3430.
14. Shteynberg D., Nesvizhskii A. I., Moritz R. L., Deutsch E. W. Combining results of multiple search engines in proteomics // Molecular & Cellular Proteomics. 2013. Vol. 12, no. 9. P. 2383–2393.
15. Branca R. M., Orre L. M., Johansson H. J. et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics // Nature methods. 2014. Vol. 11, no. 1. P. 59–62.
16. Ning K., Fermin D., Nesvizhskii A. I. Computational analysis of unassigned high--quality MS/MS spectra in proteomic data sets // Proteomics. 2010. Vol. 10, no. 14. P. 2712–2718.
17. Helmy M., Sugiyama N., Tomita M., Ishihama Y. Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics // Genes to Cells. 2012. Vol. 17, no. 8. P. 633–644.
18. Rutherford K., Parkhill J., Crook J. et al. Artemis: sequence visualization and annotation // Bioinformatics. 2000. Vol. 16, no. 10. P. 944–945.
19. Chambers M. C., Maclean B., Burke R. et al. A cross-platform toolkit for mass spectrometry and proteomics // Nature biotechnology. 2012. Vol. 30, no. 10. P. 918–920.
20. Cottrell J. S., London U. Probability-based protein identification by searching sequence databases using mass spectrometry data // electrophoresis. 1999. Vol. 20, no. 18. P. 3551–3567.
21. Fenyö D., Beavis R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes // Analytical chemistry. 2003. Vol. 75, no. 4. P. 768–774.