

Московский физико-технический институт (государственный университет)

Факультет биологической и медицинской физики

Кафедра кафедры молекулярной и трансляционной медицины

Диссертация допущена к защите

зав. кафедрой

_____ Лазарев В.Н.

«_____» _____ 2017 г.

**Выпускная квалификационная работа
на соискание степени
МАГИСТРА**

**Тема: Количественный протеоеномный
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	4
2.	Введение	5
3.	Обзор литературы	6
3.1.	Mycobacterium tuberculosis	6
3.2.	Применение масс-спектрометрии в протеомике	6
3.3.	Orbitrap	6
3.4.	Протеогеномика	6
	Типы пептидов, идентифицируемых при протеогеномном ис- следовании	7
	Подходы к созданию поисковых баз	7
	Поиск новых генов и корректировка рамок	8
	Причины приводящие к неточности аннотации	8
4.	Материалы и методы	9
4.1.	Получение бактерий	9
4.2.	Проведение масс-спектрометрического эксперимента	9
4.3.	Контроль качества	9
4.4.	Создание поисковых баз	9
4.5.	Идентификация пептидов и белков	9
4.6.	Протеогеномика <i>W-148</i>	10
	Идентификация новых белков	10
	Уточнение N-концов	10
4.7.	Сравнение идентификаций против <i>W-148</i> и <i>H37Rv</i>	10
	Поиск новых генов	10
	Уточнение N-концов	10
	Анализ SAP	10
5.	Результаты и обсуждение	11
5.1.	Протеогеномика <i>W-148</i>	11
	Идентификация	11
	Новые гены и их валидация	11
	Уточнение N-концов	11

5.2.	Сравнение идентификаций против <i>W-148</i> и <i>H37Rv</i>	11
	Новые гены и их валидация	11
	Уточнение N-концов	11
	Валидация SAP	11
6.	Выводы	12
	Список литературы	13

1. Список сокращений

GSSP - Genome Search Specific Peptides. Это пептиды, идентифицируемые при поиске против геномной базы, и не идентифицируемые при поиске 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 отличие 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 прекурсоров [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 (находясь между аннотированными генами) и внутрегенные (находятся полностью или частично в областях, где содержится аннотированный ген). Внутрегенные можно разделить на 1. находящиеся в белоккодирующих 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 экзонам, и, как правило, меньше 20% относятся к экзон-экзон участкам. Новые пептиды, не идентифицируемые против какой-либо базы, могут находиться в неаннотированных участках генома, быть результатами одно-аминокислотной замены (SAP), находится в нетранслируемых регионах (3' или 5' UTR) или интронах, являясь результатом альтернативного сплайсинга [6].

Подходы к созданию поисковых баз

Идентификация пептидов 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 увеличение 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 экзон-экзон 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 Ensembl 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 референсной базе могут быть дополнены последовательностями, являющимися 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

Поиск новых генов и корректировка рамок

Причины приводящие к неточности аннотации

4. Материалы и методы

4.1. Получение бактерий

4.2. Проведение масс-спектрометрического эксперимента

4.3. Контроль качества

4.4. Создание поисковых баз

В работе использовалось 2 типа баз: белковая и геномная. Белковая база - аннотированные последовательности, для данного штамма. Геномная - 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 (NCBI Reference Sequence: NZ_CP012090.1, 4020 аминокислотных последовательностей для W-148 и). Геномные базы были получены в результате 6 рамочного транслирования от стоп- до стоп-кадрона геномов штаммов *M.tuberculosis* W-148 и *M.tuberculosis* H37Rv, используя программу Artemis версия 16.0.0 [13]. При транслировании использовалась 11 трансляционная таблица NCBI. Минимальная длина рамки была установлена в 100 нуклеиновых кислот. К каждой базе были добавлены последовательности 26 контаминантных белков (кератины, альбумины, трипин).

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 [14]. Идентификация проходила против двух белковых и двух геномных баз с использованием Mascot Search Engine version 2.5.1 [15]. Параметры поиска были следующими: триптические пептиды, не более двух пропущенных сайтов трипсинолиза, ошибка массы прекурсера 20 ppm, ошибка массы фрагментов 0.05 Да, заряды прекурсера 2+, 3+, 4+. Oxidation(M), Carbamidomethylation(C) and Deamidated(NQ) были установлены как возможные модификации пептидов. Для подсчета FDR и порогового скоринга использовался поиск против decoy-базы, полученной в результате реверса исходной базы. FDR был выбран на уровне 5%. Пептид считался идентифицированным, если его скор выше порогового скоринга и ранг равен единице. Белок считался

идентифицированным, если для него нашлось два и более уникальных 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*.

Идентификация новых белков

Рассматривались ORF, в которых было идентифицировано два и более уникальных пептидов.

Уточнение N-концов

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

Поиск новых генов

Уточнение N-концов

Анализ SAP

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. Rutherford K., Parkhill J., Crook J. et al. Artemis: sequence visualization and annotation // Bioinformatics. 2000. Vol. 16, no. 10. P. 944–945.
14. 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.
15. 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.