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-спекртрометрии 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 поулченых MS/MS спектров. В отличае от геномных технологий, вроде ДНК или РНК секвенирования, где происхоид 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]. 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 аннотированым 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 подходов являтеся 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 спектрво 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].

Подходы к созданию баз

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. 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 (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-каднона геномов штаммов *M.tuberculosis W-148* и *M.tuberculosis H37Rv*, используя программу Artemis версия 16.0.0 [7]. При транслировании 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).

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 [8]. Идентификация проходила против двух белковых и двух геномных баз с использованием Mascot Search Engine version 2.5.1 [9]. 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 устанолвены как возможнные 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 еденице. Белок считался

идентифицированным, если для него нашлось два и более уникальых 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 идентифицированно два и более уникальных пептидов.

Уточнение N-концов

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

Поиск новый 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. Rutherford K., Parkhill J., Crook J. et al. Artemis: sequence visualization and annotation // Bioinformatics. 2000. Vol. 16, no. 10. P. 944–945.
8. 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.
9. 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.