Поиск альтернативы MaxQuant для анализа протеомных данных (масс-спектрометрия без использования метки)

Полина Дроздова

Руководитель: Дарья Бедулина (НИИ биологии ИГУ)

10 сентября 2015 г.

Введение

Сегодня количественная протеомика на основе данных масс-спектрометрии гидролизированных белков — это самый популярный метод изучения протеома живых клеток.

Одним из наиболее распространённых решений для анализа таких данных является программа MaxQuant [1], представляющая собой многофункциональную платформу для идентификации белков по масс-спектрам с использованием как общедоступных, так и собственных баз данных, и количественного анализа белков. Программа проста в использовании и проверена временем, однако у неё есть такие серьёзные недостатки, как запуск только в операционной системе Windows и работа в режиме GUI. Целью данного проекта является поиск альтернатив MaxQuant и сравнение их применимости к решению с использованием конкретных данных протеогеномного проекта «Протеомные механизмы стресс-ответа эндемичной байкальской амфиподы *Eulimnogammarus cyaneus*».

В задачи проекта входит:

1. анализ алгоритмов, применяемых на платформе MaxQuant, и поиск существующих альтернатив;
2. сравнение существующих альтернатив и выбор;
3. разработка пайплайна, позволяющего автоматизировать анализ больших объёмов подобных данных в будущем.

В качестве тестируемых данных использовали пептидные спектры триптических гидролизатов тотального белка амфипод, предварительно фракционированного по молекулярной массе методом 1D PAGE, полученные методом nano-HPLC/nano-ESI-MS/MS на приборе LTQ Orbitrap (Thermo Scientific), поскольку именно такие данные планируется обрабатывать в дальнейшем.

Разнообразие существующих решений

Для составления списка программ использовали обзорные статьи и поиск в сети Интернет (см. таблицу; курсивом выделены программы, которые предполагается проверять на следующем этапе).

Программа	Платформа	Описан в обзоре	Достоинства	Проблемы
APP	*WML*	*Свободный поиск*	*Очень подробный скрипт для установки.*	
Census	WML	[5]	Поддерживается (последняя версия — 2014 г.)	Не смогла установить
Corra	L	[5], [7], [4]		Последняя версия — 2010 г.
msBID	WL	[4]	Предназначен для работы с данными без метки.	Много зависимостей: JDK, R, Perl. Запустить не смогла: запутанная документация.
msCompare	L	[7]		Сложная процедура создания файла конфигурации.
msInspect	WML	[5], [4]		Специфическая сфера использования. Последняя версия — 2010 г.
mzMine2	WML	[5]	Работает: читает mzML, извлекает пики и делает нормализацию (по крайней мере, в режиме GUI).	reading native raw is windows only feature; the necessity to adjust parameters manually; the WOST thing—doesn't seem to work with sequence-based libraries; was designed to work as GUI
OpenMS/TOPP	*WML*	*[5], [7], [4]*	*Вероятно, в будущем — один из принятых стандартов.*	
PeptideShaker	*WLM*	*Свободный поиск*	*Версия 1.0.1 — август 2015 г.*	
Proteios	WML	[7]		Сложная установка сервера MySQL.
Proteomatic	WML	Свободный поиск		Не смогла запустить. Последняя версия — 2012 г.

ProteoSuite	WML	Свободный поиск		Документация явно не дописана, последняя версия — 0.35.
PVIEW	?	[4], [5]		Не смогла установить зависимости + документация недостаточно подробная
RforProteomics	WLM	Свободный поиск		R => медленно. Авторы честно предупреждают, что поиск лучше делать с помощью сторонних инструментов.
Sashimi	WML	Свободный поиск		
CPFP	*WML*	*Свободный поиск*	*Очень подробный скрипт для установки и документация вообще.*	*Могут возникнуть проблемы с версией Perl.*
SuperHirn	ML	[5], [4]		Последняя версия — 2009 г.
TPP	*WML*	*Свободный поиск*	*Вероятно, в будущем — один из принятых стандартов.*	*Версия 4.8.0 работает странно.*
Viper	W	[5]		Работает только под Windows.
PEPPeR	W	[4]		Заявлен как бесплатный инструмент, но что-то не похоже.
MSight	W	[4]		Похоже, работает только под Windows. Давно не поддерживается.
ProtQuant	W	[4]		Работает только под Windows.
MSQuant	W	[4]		Работает только под Windows.
IDEAL-Q	W	[4]		Работает только под Windows.
APEX	?	[4]		Сайт не работает
ProteinQuant	?	[4]		Сайт не работает

1 Форматы файлов и перевод из одного формата в другой

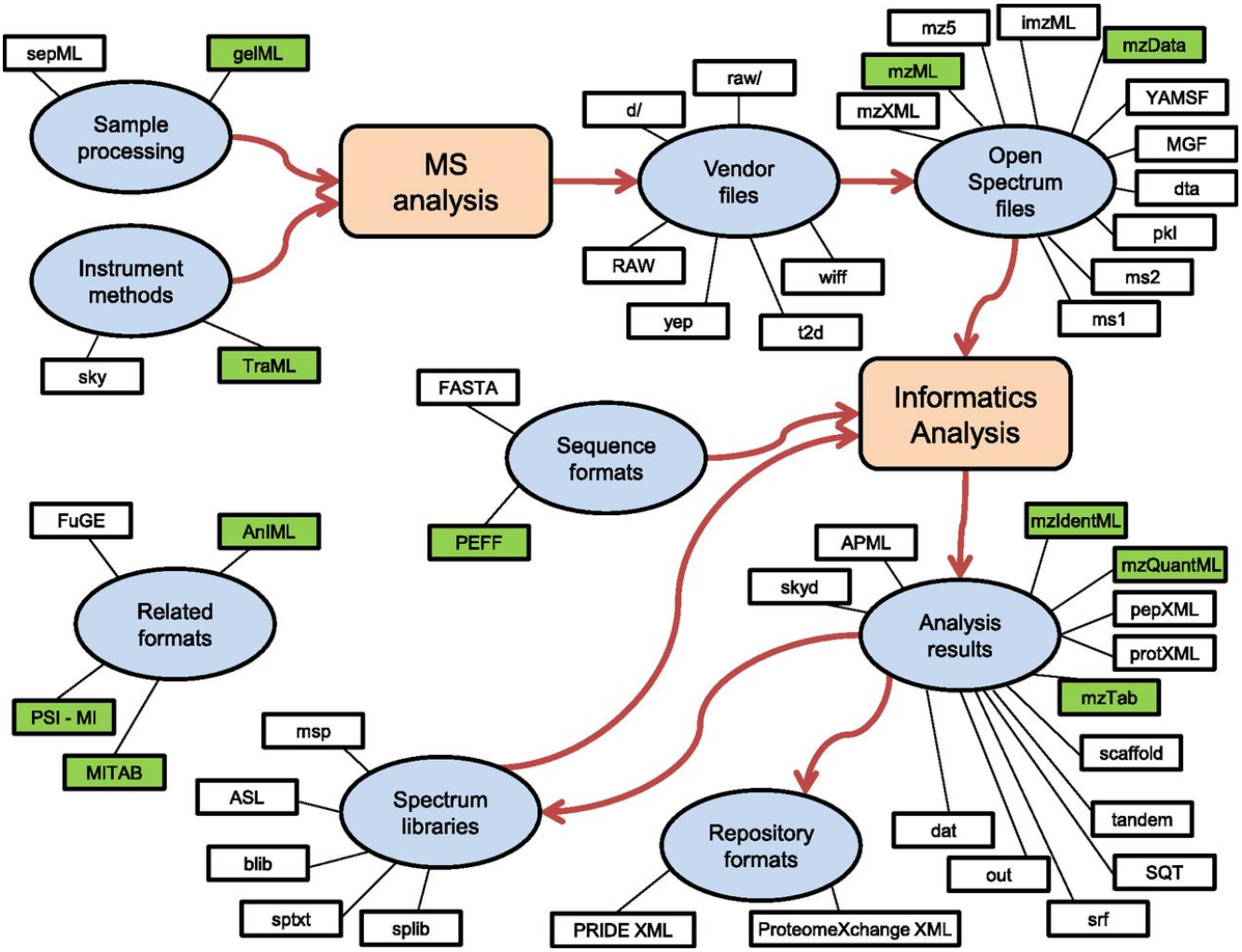

Рисунок 1. Схема, описывающая разнообразие и взаимоотношения между существующими форматами протеомных данных. Источник: [2].

Одной из самых распространённых проблем в обработке данных, в том числе данных протеомных исследований — разнообразие форматов их записи. Производители масс-спектрометров разрабатывают собственные проприетарные форматы данных. Такие форматы данных часто являются бинарными, что позволяет экономить память, но затрудняет чтение данных. Ещё одной проблемой является то, что новые версии проприетарных пакетов для работы с данными не работают с данными, записанными более старыми версиями программ. Более удобным представляется использование единого открытого формата, не зависящего от использованного прибора. Среди открытых форматов для записи сырых данных следует упомянуть mzML и mzMXL [2]. Для перевода данных из формата RAW существуют несколько программ.

1.1 ReAdW и msconvert

Trans Proteomics Pipeline (см. ниже) предлагает несколько библиотек для перевода данных из формата RAW в открытые форматы, однако все эти библиотеки требуют использования Windows.

1.2 unfinnigan

Автор проекта unfinnigan [http://code.google.com/p/unfinnigan/] просит рассматривать его как инструмент для просмотра и анализа содержимого файлов в формате RAW, а не как инструмент для изменения формата, но следует отметить, что unfinnigan успешно выполняет работу по переводу формата из RAW в mzML.

2 MaxQuant

MaxQuant использует алгоритмы подсчёта, которые можно осуществить и с помощью других программ. Основной особенностью этой программы является собственный алгоритм поиска, Andromeda.

Как уже сказано выше, недостатками MaxQuant, мешающими применять эту программу для обработки большого объёма данных, являются запуск исключительно на основе операционной системы Windows и работа в графическом режиме. Приемлемым вариантом запуска MaxQuant из командной строки является mq-run [http://mqrun.readthedocs.org/en/latest/], однако это решение не избавляет от необходимости работы на базе Windows. MaxQuant объявил о возможном появлении версии для Linux [http://141.61.102.17/maxquant_doku/], но этого не произошло.

3 Corra

Не получилось запустить. Кроме того, последняя версия датирована 2010 годом, и не похоже, что проект поддерживается.

4 msBid

Много трудностей с установкой, в основном связанных со старой версией Perl. Последняя версия программы вышла в 2008 году.

5 msCompare

Последняя версия программы вышла в 2010 году. Кроме того, документация не помогает разобраться в созании файла конфигурации в xml-подобном формате.

6 mzMine2

Удалось запустить в графическом режиме, однако потоковый режим оказался довольно неудобным.

7 PeptideShaker

Очень новый проект (27 августа 2015 г. наконец обновлён до версии 1.0) и очень подробная документация позволяют надеяться на успешное использование программы. В том числе в документации подробно описана установка программы, и она действительно не вызвала затруднений. Пример использования программы (к сожалению, также в графическом режиме) есть на странице http://compomics.com/bioinformatics-for-proteomics/.

8 ProteoSuite

Эта программа опубликована [3], но проект явно находится в стадии активной разработки [https://bitbucket.org/slurmclassic/proteosuite], и пока разобраться в документации довольно сложно.

9 PVIEW

Проект очень давно не обновлялся. Кроме того, не удалось установить часть зависимостей даже для старой версии.

10 RforProteomics

Новый активно разрабатываемый пакет, но скорости работы R может оказаться недостаточно для обработки большого объёма данных. Производители честно признаются в том, что для идентификации пептидов лучше использовать сторонние программы [6].

11 TPP и обёртки для TPP

TPP — один из хороших кандидатов на роль, поскольку эта программа одновременно достаточно давно разработана и активно поддерживается, однако эту её достаточно сложно как установить, так и использовать.

Вероятно, по этой причине написаны несколько программ, основанных на TPP, но более удобных в обращении.

11.1 TPP

TPP может быть использован и в графическом режиме, и в консольном режиме, однако руководства предполагают использование именно в графическом режиме.

11.2 APP

APP, automated proteomics pipeline, — это новая программа, работающая в том числе на основе TPP. Работающие инструкции по установке также внушают оптимизм относительно возможного использования.

11.3 CPFP

К недостаткам такого выбора можно отнести то, что

11.4 Sashimi

Sashimi включает TPP и некоторые утилиты для удобства работы, однако проблемы, возникшиеся при использовании этой программы, сводятся к проблемам использования собственно TPP.

12 TOPP/OpenMS

Ещё один достаточно часто используемый набор программ,
Относительно сложная процедура установки, но, вероятно, проблемы решаемы.

Заключение

Наш анализ позволил показать, что:

1. Программное обеспечение для обработки протеомных данных на сегодня разработано ощутимо хуже, чем для геномных, однако область активно развивается.
2. Свободный поиск в сети Интернет оказался намного более продуктивным способом изучения разнообразия программ, чем чтение обзорных статей (что, вероятно, связано с п. 1: наиболее новые проекты не успевают попасть в обзорные статьи, а многие проекты уже оказываются заброшенными).
3. Наиболее многообещающие программы — TOPP/OpenMS, TPP и PeptideShaker (после выхода версии 1.0). Кроме того, имеет смысл следить за продвижением ProteoSuite.

Список литературы

1. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26,** 1367—72. ISSN: 1546-1696 (дек. 2008).
2. Deutsch, E. W. File formats commonly used in mass spectrometry proteomics. *Molecular & cellular proteomics : MCP* **11,** 1612—21. ISSN: 1535-9484 (дек. 2012).
3. Gonzalez-Galarza, F. F. *и др.* A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. *Omics : a journal of integrative biology* **16,** 431—42. ISSN: 1557-8100 (сент. 2012).
4. Lemeer, S., Hahne, H., Pachl, F. & Kuster, B. Software tools for MS-based quantitative proteomics: a brief overview. *Methods in molecular biology (Clifton, N.J.)* **893,** 489—99. ISSN: 1940-6029 (янв. 2012).
5. Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for label-free peptide quantification. *Molecular & cellular proteomics : MCP* **12,** 549—56. ISSN: 1535-9484 (март 2013).
6. Gatto, L. & Christoforou, A. Using R and Bioconductor for proteomics data analysis. *Biochimica et biophysica acta* **1844,** 42—51. ISSN: 0006-3002 (янв. 2014).
7. Sandin, M., Teleman, J., Malmström, J. & Levander, F. Data processing methods and quality control strategies for label-free LC-MS protein quantification. *Biochimica et biophysica acta* **1844,** 29—41. ISSN: 0006-3002 (янв. 2014).