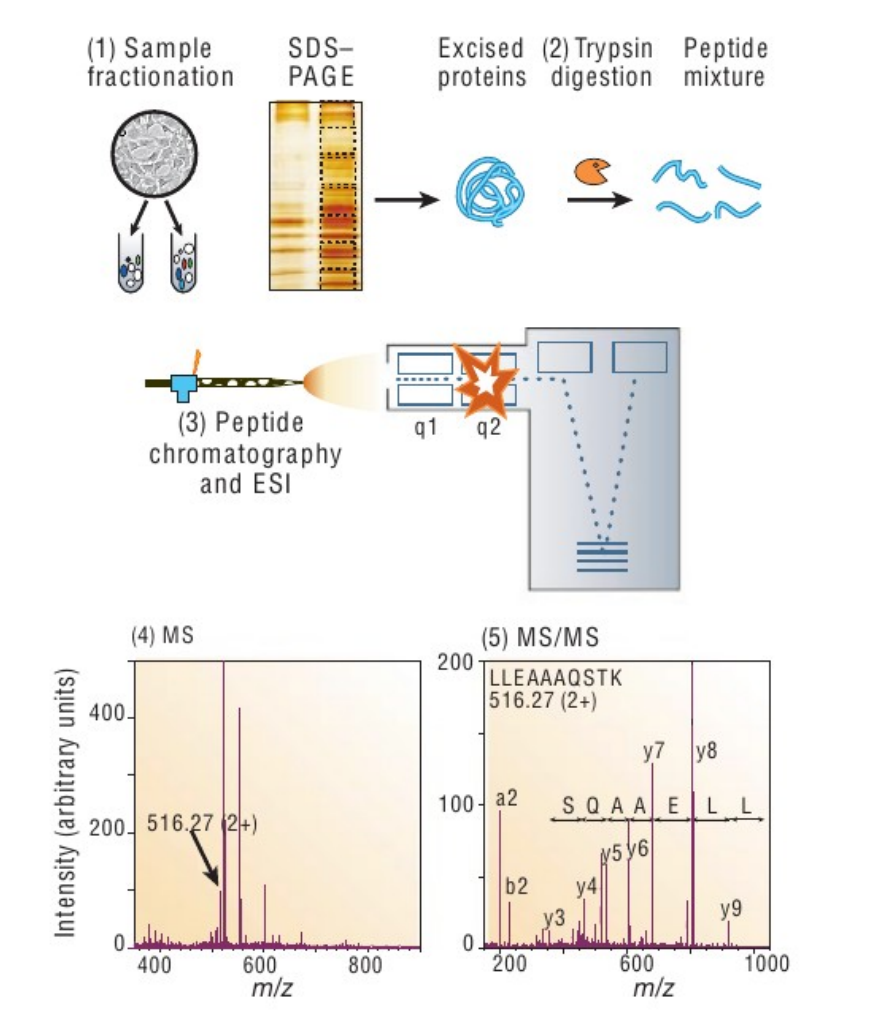
Открытые инструменты для работы с протеомными данными на языке программирования Python

Л.И. Левицкий, М.В. Иванов, А.А. Голобородько, М.В. Горшков

Институт энергетических проблем химической физики им. Тальрозе РАН
Московский физико-технический институт (государственный университет)

Схема протеомного эксперимента (bottom-up)



- Пробоподготовка, фракционирование
- Ферментативный гидролиз
- Хроматографическое разделение (LC), ионизация (ESI)
- Тандемная масс-спектрометрия (MS/MS)

Ruedi Aebersold & Matthias Mann, “Mass spectrometry-based proteomics”. *Nature, 422 (6928), pp. 198-207 (2003).*

Pyteomics: структура и функциональность

- Аминокислотная последовательность (`parser`)
  - modX-нотация
  - парсинг последовательностей
  - аминокислотный состав
  - модификации
- Предсказание свойств пептидов (белков)
  - масса, m/z, изотопные распределения (`mass`)
  - хроматографические времена (`achrom`, `biolccc*`)
  - изоэлектрическая точка, заряд в растворе (`electrochem`)
- Работа с данными в стандартных форматах
  - данные LC-MS/MS (`mzml`, `mgf`)
  - вывод поисковых машин (`pepxml`, `mzid`, `tandem`)
  - базы данных белковых последовательностей (`fasta`)
- Визуализация (`pylab_aux`)
- Вспомогательные функции (`auxiliary`)

Работа с последовательностями: `pyteomics.parser`

modX notation: “H-oMYPEpTIDE-OH”

- Длина последовательности
- Парсинг (*str* → *list*)
- Произвольные модификации (постоянные, потенциальные)
- Аминокислотный состав
- Гидролиз *in silico* (произвольные правила расщепления, заготовки для распространённых ферментов , “недорезы”)

Свойства белков и пептидов

`pyteomics.mass`

- Моноизотопная масса (m/z)
  - По формуле
  - По аминокислотной последовательности
  - По аминокислотному составу
- Масса (m/z) конкретного изотопного состояния
- Вероятность изотопного состояния
- Наиболее вероятное изотопное состояние
- Арифметика атомных и изотопных составов

`pyteomics.electrochem`

- Изоэлектрическая точка (pI)
- Заряд при данном pH

Свойства белков и пептидов

`pyteomics.biolccc`

- Предсказание времени удерживания
- Градиентный и изократический режим
- Вычисление адсорбционных свойств пептидов
- Изменяемые условия разделения
- Модели для коротких пептидов и длинных белков
- Произвольный профиль градиента

Модель BioLCCC: ***Gorshkov, A. V., et al., Analytical chemistry, 2006***

`pyteomics.achrom`

“Аддитивная модель”:

$$RT = (1 + m\ lnN) \sum_{i=1}^{i=N} RC_i n_i + RT_0$$

- Нахождение коэффициентов удержания (калибровка)
- Вычисление времени удерживания
- Логарифмическая коррекция на длину

Meek, J. L. PNAS, 1980

Mant, C. T.; Zhou, N. E.; Hodges, R. S. Journal of Chromatography A, 1989

Работа с данными

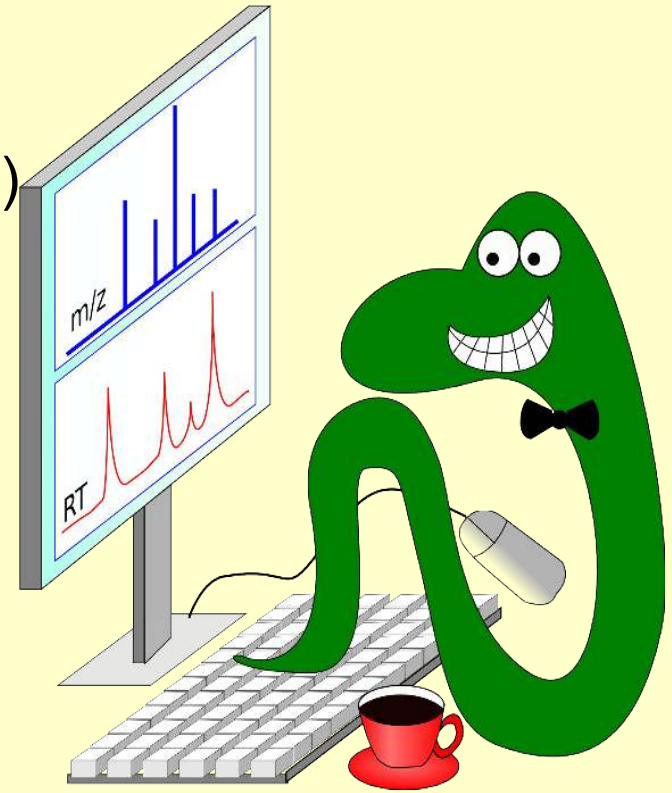
- Экспериментальные данные (**mzML, MGF**)
  - m/z, I, z, RT…
  - создание файлов в формате MGF
- Результаты идентификации (**pepXML, mzIdentML, TandemXML**)
  - идентифицированные последовательности
  - “скоры”
  - модификации
  - m/z, I, заряды, RT...
- Базы данных белковых последовательностей (**FASTA**)
  - последовательности и аннотации
  - создание файлов FASTA
  - генерация “декойных” баз данных

Ссылки

- **Пакет на PyPI:** http://pypi.python.org/pypi/pyteomics
- **Документация:** http://pythonhosted.org/pyteomics
- **Код: http://hg.theorchromo.ru/pyteomics** (BitBucket)
- **Рассылка/форум:** http://groups.google.com/group/pyteomics
- **TheorChromo Online:** http://theorchromo.ru

Распространение

- Открытый код
- Свободная лицензия (Apache 2.0)
- Кроссплатформенность (Windows, Mac, Linux)



Публикация:
A. Goloborodko, L. Levitsky, M. Ivanov et al. “Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics”. *JASMS, 24 (2), pp. 301-304. (2013)*

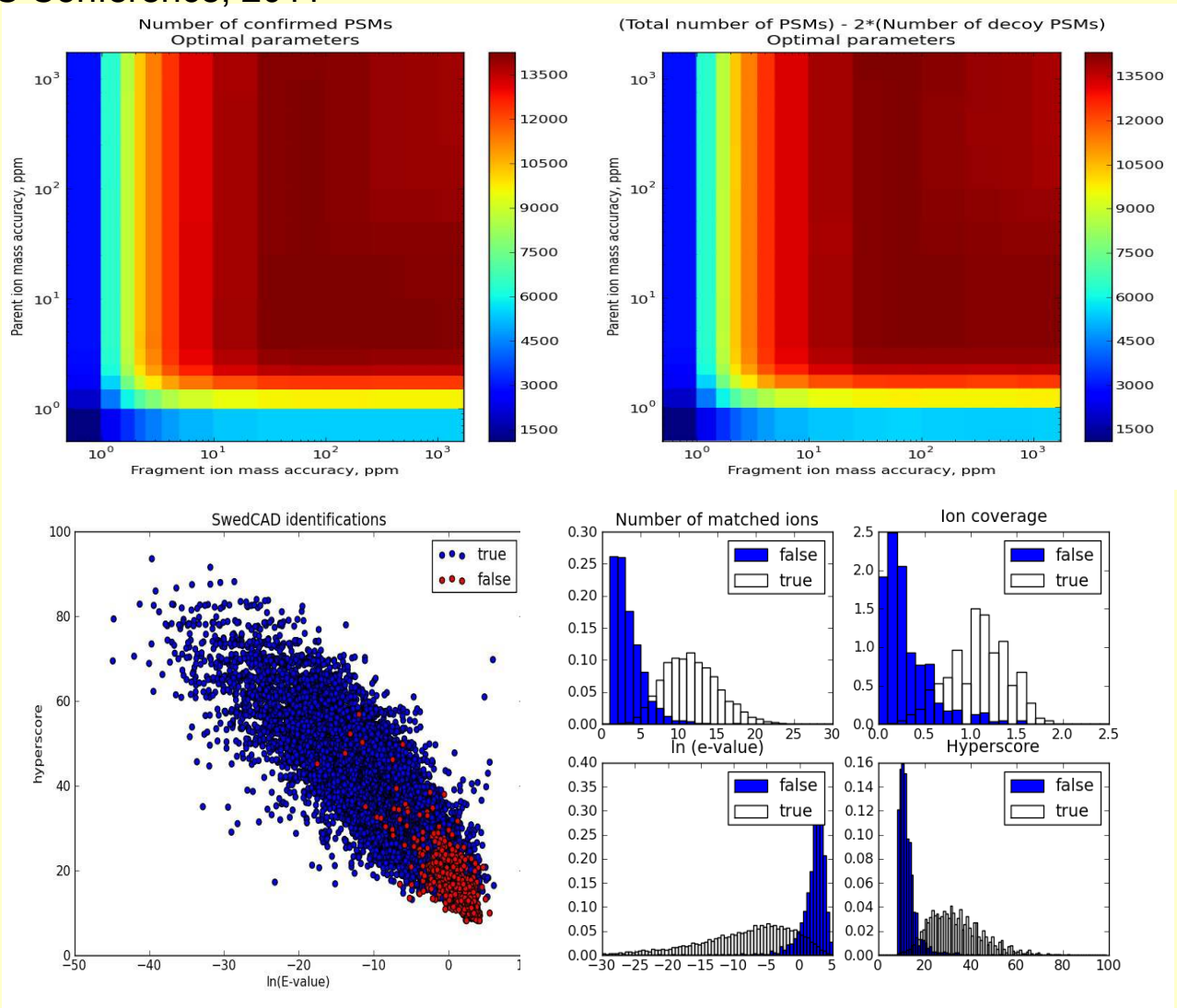
Финансирование:
7-я рамочная программа ЕС (FP7):
грант Prot-HiSPRA # 282506
РФФИ: грант # 11-04-00515


Pyteomics

Приложение 1

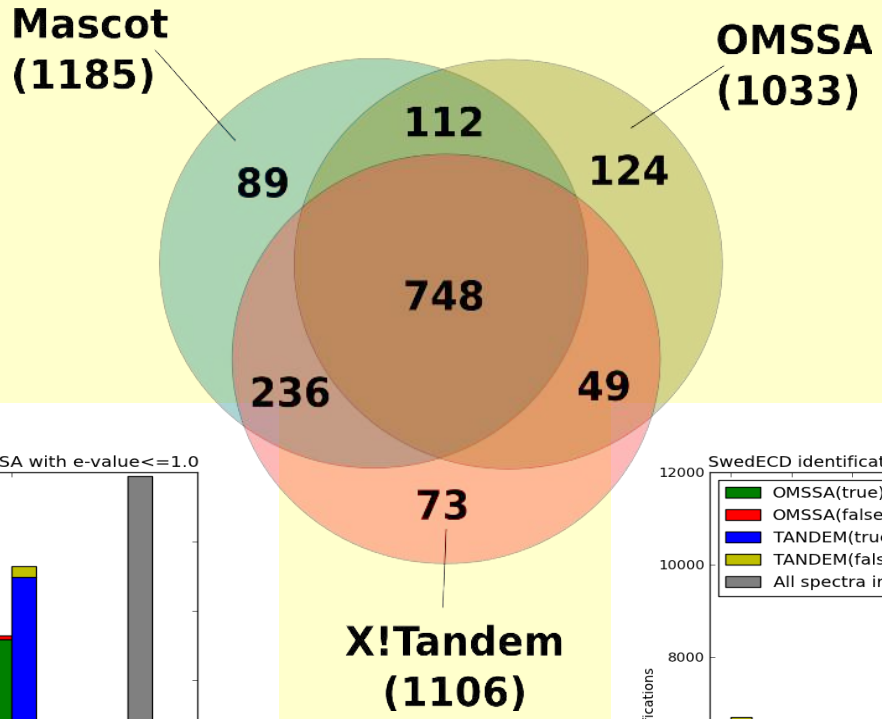
Влияние параметров поиска на количество правильных идентификаций

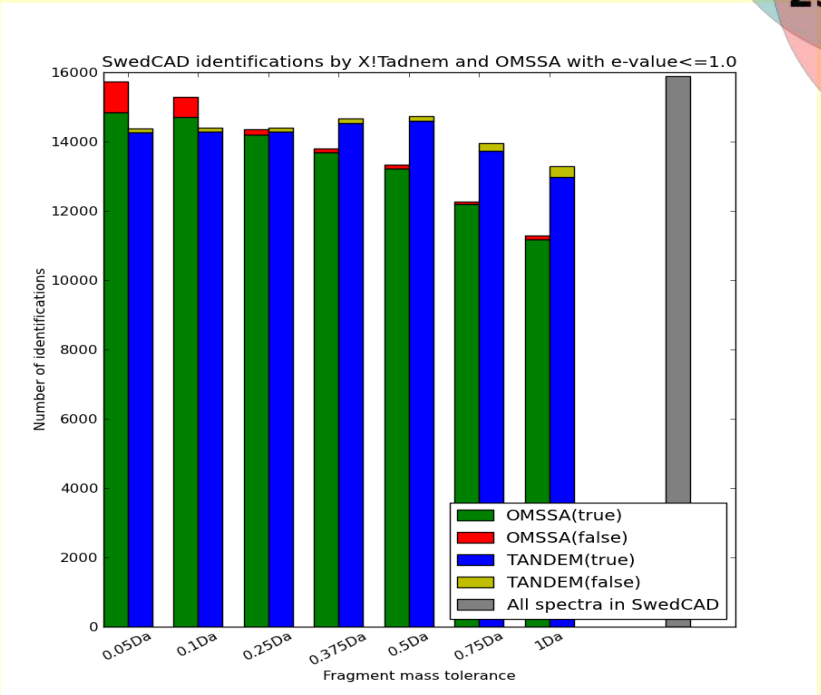
L.I. Levitsky, A.A. Goloborodko, M.V. Gorshkov, “The influence of search parameters and mass spectrometry data quality on the search engine performance in shotgun proteomics: a systematic study”. ASMS Conference, 2011

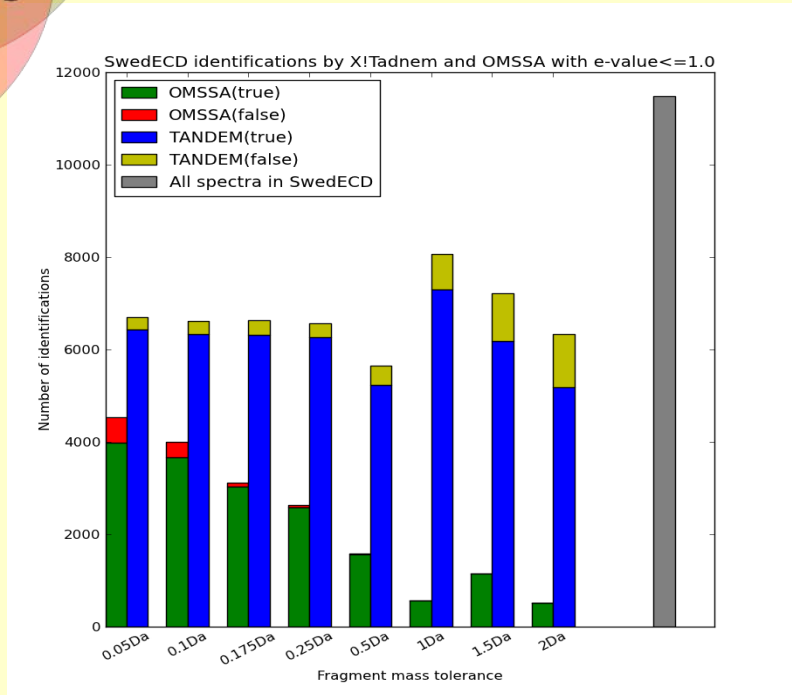


Данные: SwedCAD ~16,000 аннотированных CID-спектров (***Fälth, M., et al. Journal of Proteome Research, 2007***)

Приложение 2
Сравнительный анализ поисковых машин



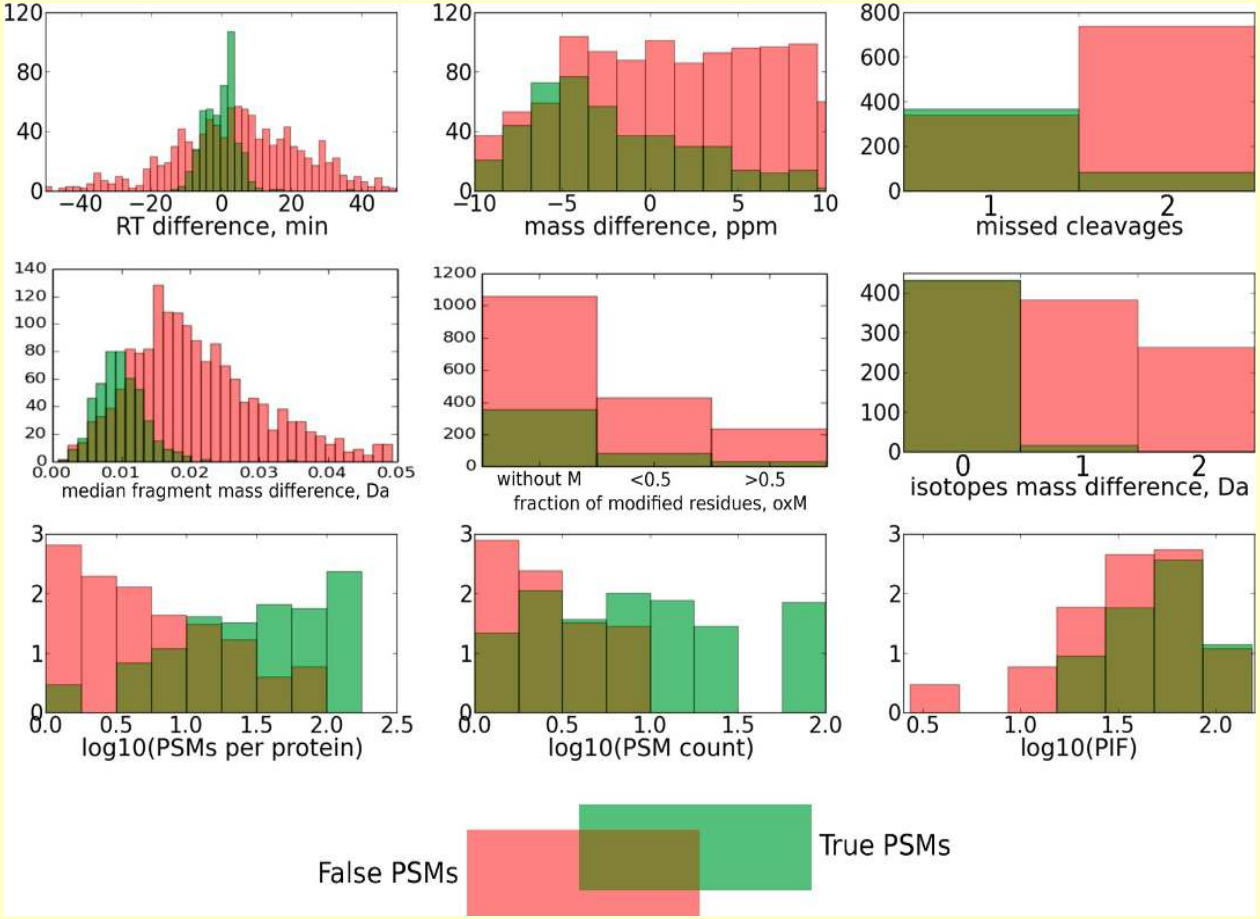


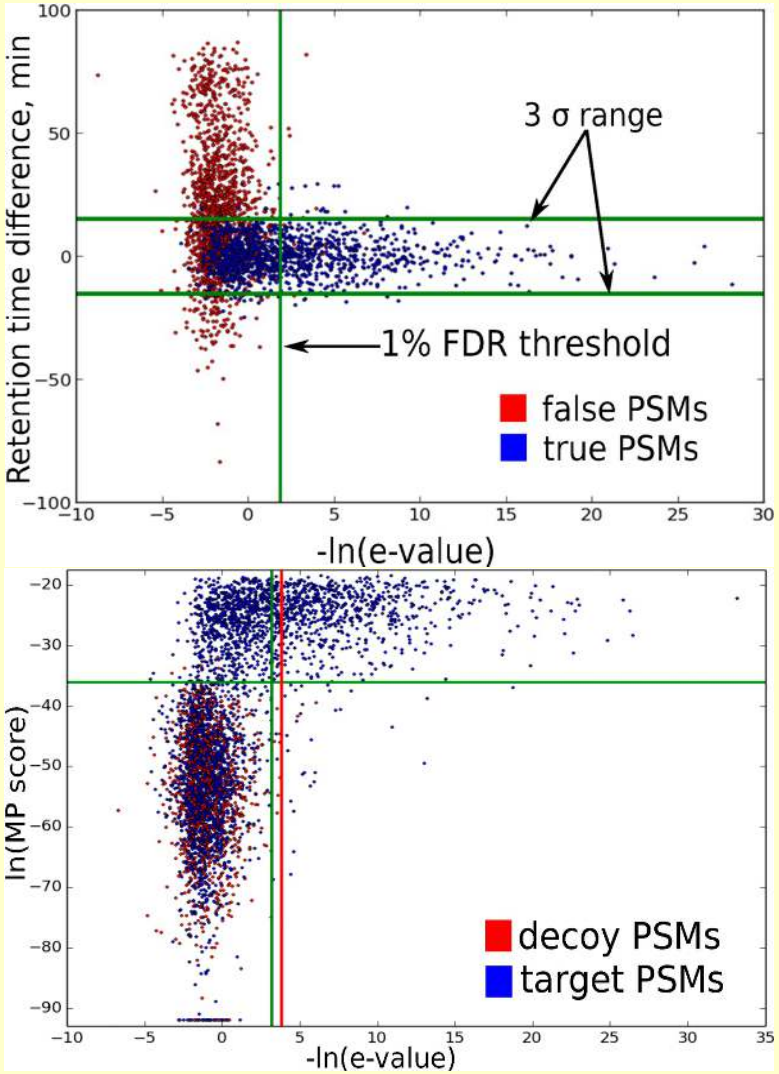


M. Ivanov, L. Levitsky, A. Goloborodko, I. Tarasova, S. Steinhammer, G. Mitulovic, M. Gorshkov, “Performance comparison of database search engines using annotated MS/MS data obtained by different dissociation techniques”, The 60th ASMS Conference on Mass Spectrometry and Applied Topics, May 19-24, 2012, Vancouver, Canada

Приложение 3

Эмпирический алгоритм валидации идентификаций MPScore

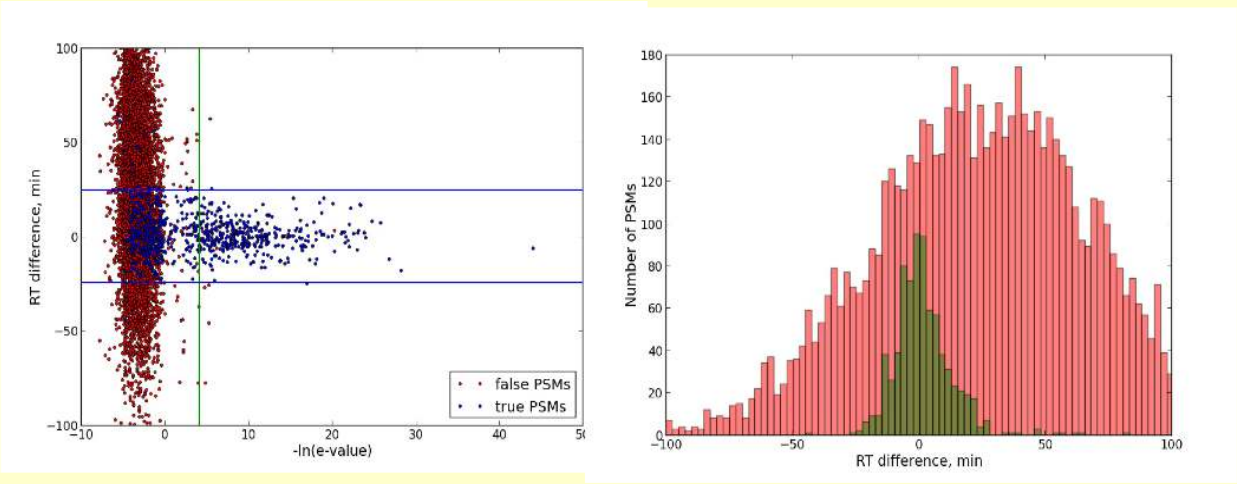


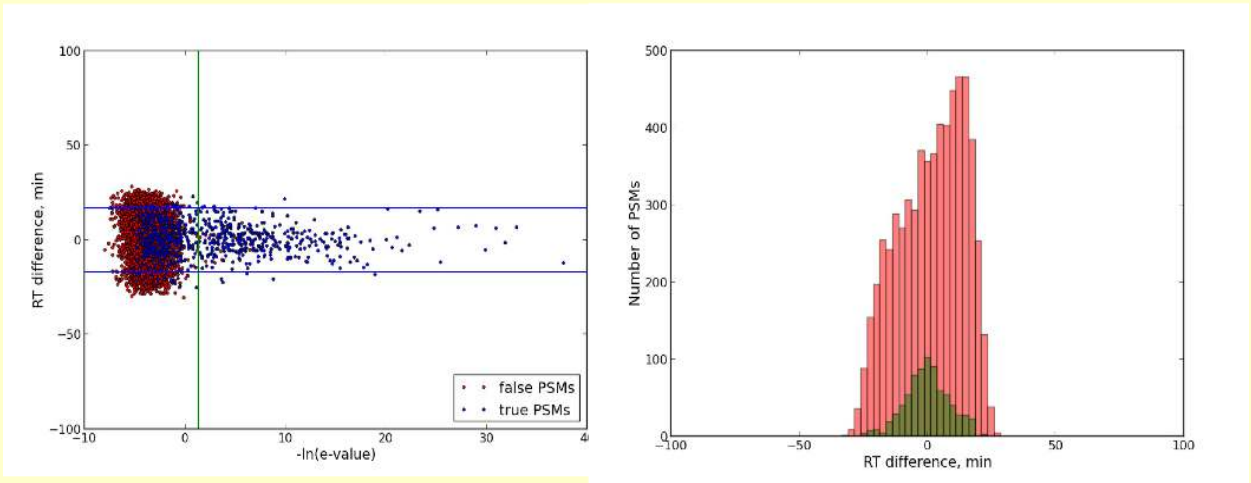


M.V. Ivanov, L.I. Levitsky, A.A. Lobas, T. Panic, U.A. Laskay, G. Mitulovic, R. Schmid, M.L. Pridatchenko, Y.O. Tsybin, and M.V. Gorshkov, “Empirical Multidimensional Space for Scoring Peptide Spectrum Matches in Shotgun Proteomics”. *Journal of Proteome Research, 13, pp. 1911-1920 (2014)*

Приложение 4

Поисковая машина IdentiPy





- Лёгкая замена скоринг-функций
- Добавление собственных СФ
- Произвольная фильтровка
- Расширяемость

L.I. Levitsky, M.V. Ivanov, A.A. Lobas, M.L. Pridatchenko, I.A. Tarasova, T. Panic, G. Mitulovic, Y.O. Tsybin, Ü.A. Laskay, and M.V. Gorshkov, An Open-Source Search Engine Utilizing a Multidimensional Scoring Algorithm for Advanced Mass Spectrometry Data Processing and Protein Identification, The 61st ASMS Conference on Mass Spectrometry and Applied Topics, June 9-13 2013, Minneapolis, MN, USA

Приложение 5

Конвертер файлов TandemXML в pepXML:

https://bitbucket.org/markmipt/pyteomics.tandem2xml

… и многое другое