

Вычислительные методы для высокопроизводительной идентификации метаболитов

Егор Щербин

Руководитель: к. т. н. Литвинов Юрий Викторович

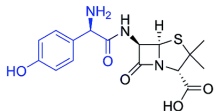
Консультант: к. ф.-м. н. Гуревич Алексей Александрович

НИУ ВШЭ СПб

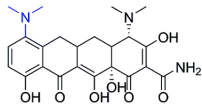
7 июня 2019 г.

Природные соединения (вторичные метаболиты) — химические соединения, производимые различными живыми организмами.

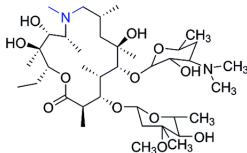
Они обладают большим потенциалом, чтобы быть антибиотиками естественного происхождения.



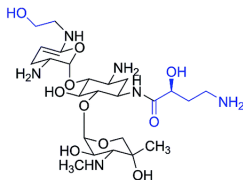
Amoxicillin



Minocycline



Azithromycin

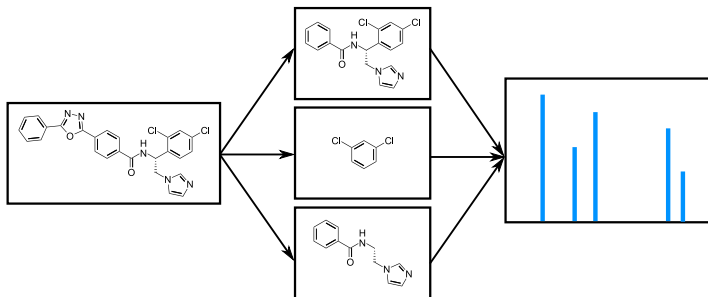


Plazomicin

Масс-спектрометрия

Тандемная масс-спектрометрия — метод химического анализа, позволяющий “взвешивать” вещества и их фрагменты.

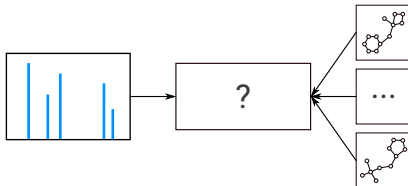
Репозиторий GNPS¹ содержит более миллиарда масс-спектров природных соединений.



¹Wang et al., Sharing and community curation of mass spectrometry data with GNPS, 2016

Идентификация масс-спектра

Задача: найти в базе химических соединений вещество, породившее данный масс-спектр.



Существующие методы:

- iSNAP² доступен только в виде веб-приложения для анализа лишь одного спектра за раз.
- CSI:FingerID³ работает слишком медленно для достаточно больших молекул (>500 Да).

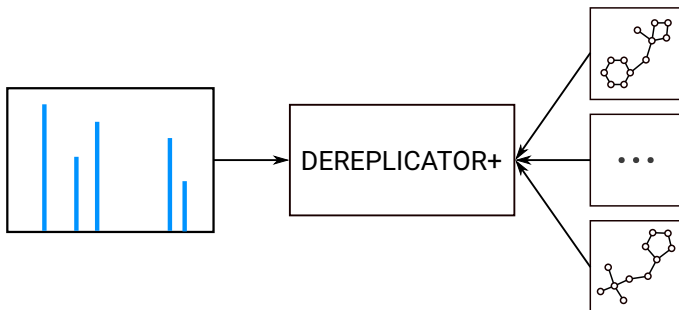
²Ibrahim et al., Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery, 2012

³Dührkop et al., Searching molecular structure databases with tandem mass spectra using CSI:FingerID, 2015

DEREPLICATOR+

DEREPLICATOR+⁴ встроен в инфраструктуру GNPS и позволяет автоматически обрабатывать поступающие наборы масс-спектров.

К сожалению, все еще недостаточно эффективен для обработки имеющегося количества данных.



⁴Mohimani et al., Dereplication of microbial metabolites through database search of mass spectra, 2018

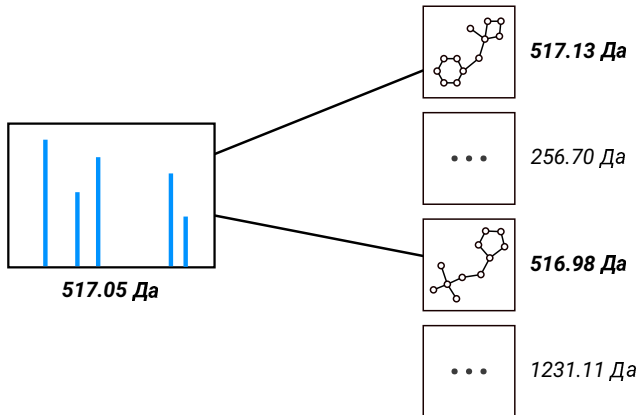
Цель: разработать методы улучшения DEREPLICATOR+, позволяющие обеспечить обработку больших объемов данных за короткое время.

Задачи:

- Изучить текущую реализацию алгоритма.
- Проанализировать производительность алгоритма.
- Предложить методы оптимизации алгоритма.
- Реализовать предложенные методы и сравнить производительность с базовой версией.

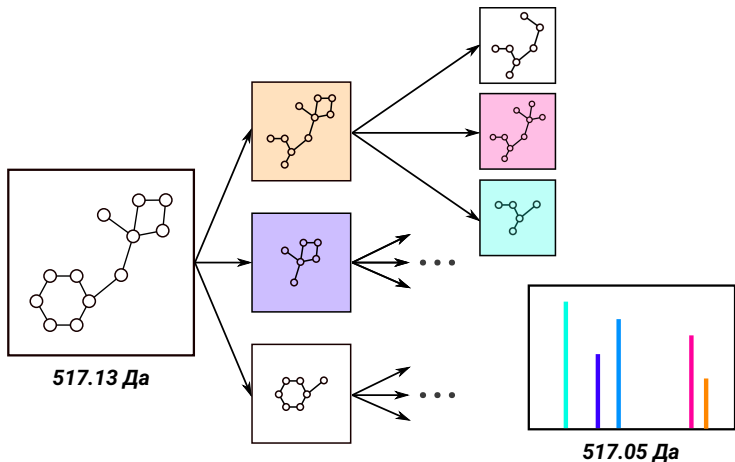
Поиск кандидатов

Спектр и молекула являются потенциальным соответствием, если у них почти совпадает масса.



Оценивание кандидатов

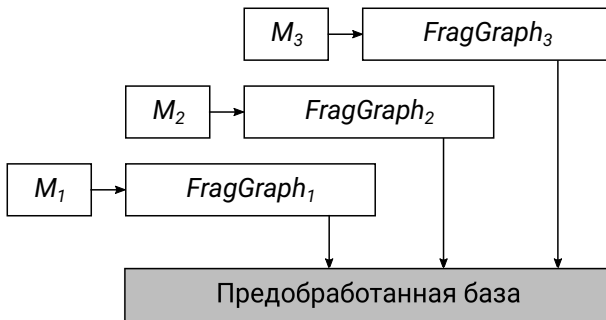
Каждое потенциальное соответствие оценивается по количеству пиков в спектре, “объясненных” графами фрагментации.



Предобработка базы

Графы фрагментации предподсчитываются для экономии времени.

Предобработанную базу необходимо на старте загрузить в память целиком.



Производительность DEREPLICATOR+ для базы AntiMarin⁵ из 60K молекул и набора данных SpectraSTREP из 234K спектров:

- Предобработка базы

Размер базы	Время	Размер файла
60908	62 ч ⁶	6.9 Гб

Вывод: использование базы размером уже на 2 порядка больше невозможно.

⁵Blunt et al, AntiMarin Database, 2006

⁶В однопоточном режиме

Производительность DEREPLICATOR+ для базы AntiMarin из 60K молекул и набора данных SpectraSTREP из 234K спектров:

- Поиск в базе

Время поиска кандидатов	Время оценивания	Кол-во кандидатов	Кол-во итоговых соответствий
0.8 мин	70 мин	487443	762

Вывод: оценивается слишком много ложноположительных кандидатов.

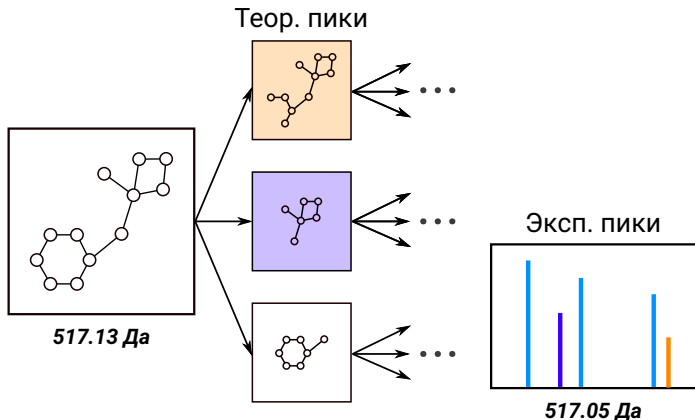
Разработаны следующие методы:

- Оптимизация хранения предобработанной базы.
- Ускоренное построение графа фрагментации.
- Строгий отбор потенциальных соответствий.

Сильный критерий для кандидатов

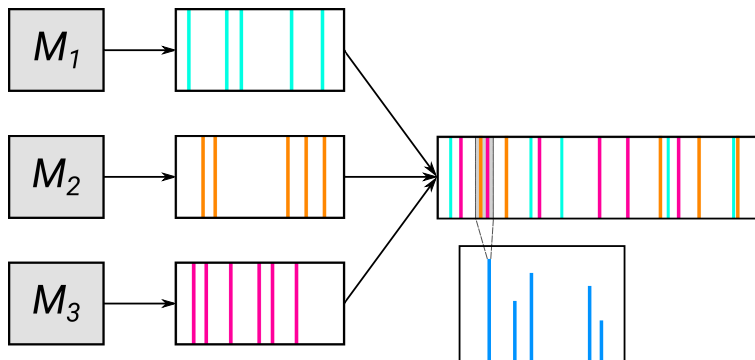
Отбираем кандидатов по количеству пиков, найденных на первом уровне фрагментации.

Критерий: количество совпадающих эксп. и теор. пиков
пиков $\#FirstLevelPeaks \geq T$.



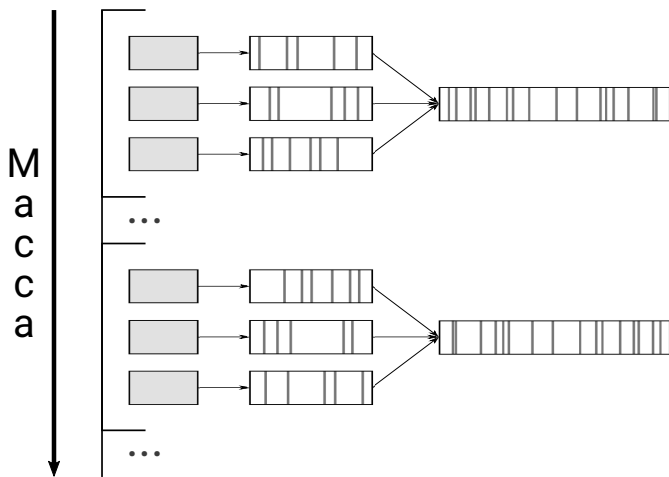
Индекс по первому уровню фрагментации

Для этого построим индекс, в котором будут храниться нужные теор. пики для рассматриваемых молекул.



Разбиение по массе

Разбиваем базу на блоки по массе, и для каждого блока строим отдельный индекс.



Тестирование производилось на узле вычислительного кластера с 128 Гб доступной памяти.

DEREPLICATOR+ поддерживает параллельную многопоточную обработку данных, но для чистоты анализа все результаты приведены для запуска в однопоточном режиме.

- AntiMarin (60908 молекул)

Версия	Время	Размер итогового файла	Размер индекса
Базовая	62 ч	6.9 Гб	-
Улучшенная	6.6 ч	6.9 Гб	4.3 Мб

Результаты. Предобработка

- AntiMarin+HMDB⁷+DNP⁸ (357554 молекул)

Версия	Время	Размер итогового файла	Размер индекса
Улучшенная	21.3 ч	30 Гб	16 Мб

- PubChem10M⁹ (10016769 молекул)

Версия	Время	Размер итогового файла	Размер индекса
Улучшенная	6.6 ч	142 Гб	623 Мб

Базовая версия не использовалась из-за больших требований по времени или памяти.

⁷Wishart et al., HMDB 4.0 — The Human Metabolome Database for 2018, 2018

⁸Gozalbes et al., Small molecule databases and chemical descriptors useful in chemoinformatics: an overview, 2011

⁹Kim et al., PubChem 2019 update: improved access to chemical data, 2019

Результаты. Поиск кандидатов

Время на отбор потенциальных соответствий:

База	Версия	Время	Кол-во кандидатов
AntiMarin	Базовая	51 сек	487443
	Улучшенная	2 сек	226285
AM+HMDB+DNP	Базовая	813 сек	3030862
	Улучшенная	4 сек	1453473
PubChem10M	Базовая	28475 сек	29498264
	Улучшенная	5 сек	7662395

Обработка набора SpectraSTREP (234928 спектров) в базе AntiMarin+HMDB+DNP (357554 молекул):

Версия	Время оценивания	Кол-во кандидатов	Кол-во итоговых соответствий
Базовая	363 мин	3030862	1167
$T = 1$	185 мин	1453473	1167
$T = 2$	26.8 мин	261117	1054
$T = 3$	10.7 мин	63019	713

Изучен и проанализирован алгоритм DEREPlicATOR+, и предложены методы его оптимизации:

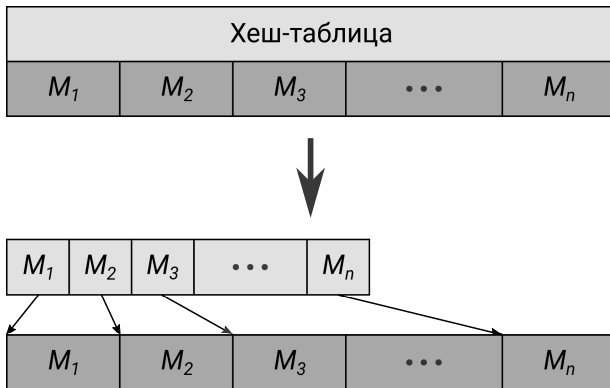
- Оптимальное хранение предобработанной базы.
- Ускоренная фрагментация молекул с помощью битовых множеств.
- Строгий отбор потенциальных соответствий с помощью индексной структуры.

Данные методы были реализованы и протестированы:

- Стала возможна работа с базами в сотни раз больше, чем раньше.
- Предобработка базы стала быстрее в 10 раз.
- Достигнуто ускорение поиска в 2 раза без потери качества, в 13 раз с потерей 10%.

Предобработка базы

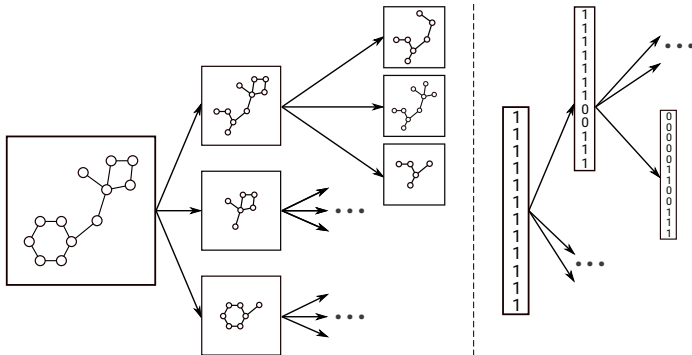
Можно добавить небольшой индекс, чтобы загружать фрагментации на лету.



Быстрая генерация фрагментаций

Построение графа фрагментации — вычислительно сложный процесс.

Его можно сильно оптимизировать, используя битовые множества для хранения структуры фрагментов.



Обработка набора SpectraSTREP (234928 спектров) в базе AntiMarin (60908 молекул):

Версия	Время оценивания	Кол-во кандидатов	Кол-во итоговых соответствий
Базовая	70 мин	487443	762
$T = 1$	38 мин	226285	762
$T = 2$	10.6 мин	55446	659
$T = 3$	4.5 мин	14281	435

Обработка набора SpectraSTREP (234928 спектров) в базе PubChem10M (10016769 молекул):

Версия	Время оценивания	Кол-во кандидатов	Кол-во итоговых соответствий
Базовая	1716 мин	29498264	2860
$T = 1$	384 мин	7662395	2860
$T = 2$	78 мин	1388791	2651
$T = 3$	18 мин	233733	2216