*ISSN 0868–5886* *НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2006, том 16, № 2, с. 92–100*

**ОРИГИНАЛЬНЫЕ СТАТЬИ**

*УДК 621.384.668.8: 577.112.6*

**© В. В. Макаров, Я. И. Лютвинский, С. К. Савельев, А. Н. Веренчиков, Н. В. Краснов**

# АЛГОРИТМ ИЗВЛЕЧЕНИЯ АНАЛИТИЧЕСКИ ЗНАЧИМОЙ ИНФОРМАЦИИ ИЗ МАСС-СПЕКТРОМЕТРИЧЕСКИХ ДАННЫХ ЭКСПЕРИМЕНТОВ ПРОТЕОМИКИ

Описан алгоритм обработки масс-спектрометрических данных с целью извлечения информации о молекулярных массах и интенсивности сигналов компонентов пробы. Алгоритм адаптирован для обработки масс-спектров многокомпонентных смесей пептидов, характерных для экспериментов протеомики, и оперирует информацией о положении и интенсивности спектральных пиков, выделенных на предварительной стадии обработки. Показано, что при сопоставимом качестве обработки алгоритм значительно превосходит коммерческие аналоги по производительности и демонстрирует зависимость времени обработки от числа спектральных пиков, характер которой близок к линейному.

## ВВЕДЕНИЕ

Исследования структуры и функций белков в живых организмах сформировали отдельное направление молекулярной биологии, получившее наименование "протеомика" [1]. Одним из основных аналитических инструментов протеомики является масс-спектрометрия [2], а наиболее актуальными масс-спектрометрическими приложениями являются задачи идентификации и секвенирования (установление последовательности аминокислот) белков, изучения пост-трансляционных модификаций, исследования количественных характеристик белка в смеси. При решении таких задач данные масс-спектрометрических экспериментов соотносятся с базами данных известных белков, и на основании этого результаты эксперимента получают биологическую интерпретацию.

Характерной особенностью исследований протеомики является высокая сложность анализируемых проб (порядка $10^5$–$10^6$ компонентов с содержанием в диапазоне от единиц фемтомолей до $10^2$ микромолей). Необходимость исследования сложных смесей обусловливает тенденцию повышения производительности анализа, что сопряжено с ростом потоков экспериментальных данных (свыше 10 МБ/с). Работа исследователя с такими массивами данных неизбежно требует использования вычислительной техники и программного обеспечения, автоматизирующего процесс обработки и интерпретации масс-спектрометрических данных. Поэтому весьма актуальной становится разработка высокопроизводительных алгоритмов для решения этих задач.

Методы масс-спектрометрического исследования белка оперируют информацией о молекулярных массах компонентов пробы и точности их определения. Эта информация является *аналитически значимой*, т. к. на ее основе производится интерпретация масс-спектрометрических данных методами биоинформатики. Непосредственное извлечение этой информации из масс-спектрометрических данных затруднено вследствие изотопного и зарядового распределения ионов компонентов пробы, формирующих сложную структуру масс-спектра.

Природное изотопное распределение, свойственное большинству химических элементов, проявляется в масс-спектре в виде серий пиков, обозначаемых термином *изотопный мультиплет* (либо *изотопный кластер*). Среди пиков изотопного мультиплета особую роль играет *моноизотопный пик*, определяющий моноизотопную массу химического соединения. При использовании источника ионов типа электроспрей наблюдается распределение ионов по зарядовому состоянию, которое проявляется в масс-спектре в виде серий из нескольких изотопных мультиплетов, относящихся к одному и тому же компоненту пробы (рис. 1). В масс-спектрах сложных смесей серии пиков ионов различных компонентов претерпевают многочисленные наложения, что крайне затрудняет интерпретацию. Для получения информации о компонентах пробы необходимо решить задачу декомпозиции масс-спектра, которая заключается в выделении и группировке фрагментов масс-спектрометрического сигнала, относящихся к изотопным мультиплетам ионов каждого компонента пробы.

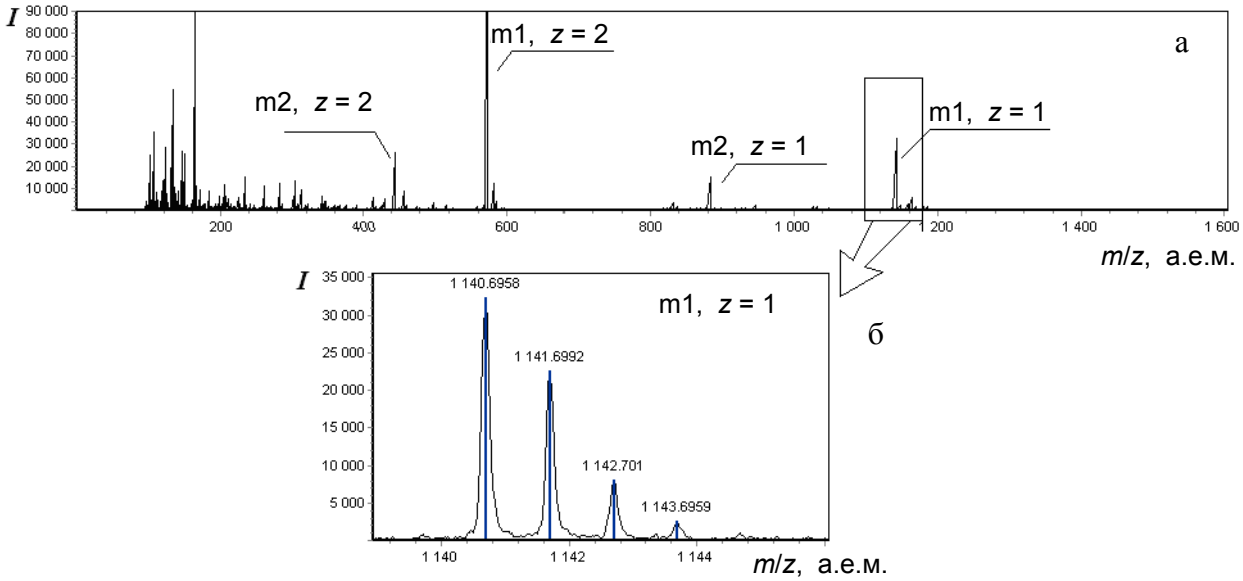


**Рис. 1.** Типичный экспериментальный масс-спектр высокого разрешения. Иллюстрация зарядового (а) и изотопного (б) распределений ионов компонентов m1 и m2

Для решения этой задачи предлагается алгоритм декомпозиции масс-спектров пептидов, оперирующий информацией о структуре сигнала и использующий специфику изотопных распределений молекул пептидов. Алгоритм отличается высокой производительностью и может быть использован в программном обеспечении перспективных масс-спектрометрических систем.

## ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ

С тех пор как метод ионизации "электроспрей" стал широко использоваться в масс-спектрометрических исследованиях биоорганических соединений, разработано большое количество методов извлечения информации о молекулярных массах веществ, экспериментально зафиксированных в масс-спектре. В период освоения этого метода широкой армией исследователей интерпретация масс-спектров производилась в ручном или полуавтоматическом режимах. Поэтому в большинстве ранних работ целью обработки является преобразование масс-спектра к виду, свободному от зарядового распределения (так называемому "нульзарядному"), более удобному для визуальной интерпретации. Один из первых методов обработки, известный под названием "деконволюция по зарядовым состояниям" (charge state deconvolution), описан в работах [3, 4]. Метод отличается простотой, однако обладает существенным недостатком: в процессе преобразования генерируется большое количество ложных пиков. Для устранения данного недостатка предложен ряд модификаций метода, в основе которых лежит использование более эффективной функции преобразования [5], а также использование принципа максимальной энтропии [6]. Данные методы характеризуются большими объемами вычислительных операций, т. к. оперируют исходным масс-спектрометрическим сигналом и сами по себе не решают задачу извлечения аналитически значимой информации. Эта задача, связанная с отсеиванием ложных пиков, возлагается на оператора либо на алгоритмы дополнительной обработки. Методы, использующие принцип максимальной энтропии, разрабатывались также в работах [7, 8]. Однако их недостатком также являются большие объемы вычислительных операций.

Интерес представляет метод, предложенный в работе [9], основанный на построении и оценке гипотез о принадлежности пиков масс-спектра к изотопным распределениям ионов различного заряда. Метод был адаптирован автором для работы с информацией о структуре сигнала, т. е. о положении и интенсивности спектральных пиков, выделенных на предыдущей стадии обработки. Алгоритм, созданный на основе предложенного метода, отличается значительно более высокой производительностью, т. к. при переходе от работы с исходным масс-спектрометрическим сигналом к работе с выделенными пиками количество

вычислительных операций сокращается более чем на порядок. Опыт успешного применения данного алгоритма для обработки масс-спектров низкого разрешения описан в работе [10]. Однако при обработке масс-спектров высокого разрешения существенным недостатком алгоритма является тот факт, что он не использует информацию о соотношении интенсивности пиков в изотопных мультиплетах. Для каждого пика масс-спектра принимается лишь одна гипотеза о зарядовом состоянии ионов, породивших пик. Тем самым упускается из рассмотрения случай, при котором пик принадлежит одновременно двум (и более) изотопным мультиплетам ионов различного заряда. Поэтому алгоритм способен корректно обрабатывать только отдельные случаи наложения изотопных мультиплетов, при которых наблюдается слияние небольшого числа пиков.

Один из путей преодоления вышеуказанных недостатков лежит через эффективное использование информации, заложенной в интенсивности пиков изотопных распределений ионов. В предлагаемом алгоритме декомпозиции масс-спектра выделение информации о компонентах пробы производится при помощи сравнения фрагментов масс-спектрометрического сигнала с так называемым *модельным изотопным мультиплетом*, который получен осреднением изотопных мультиплетов молекул реальных пептидов.

## ОПИСАНИЕ АЛГОРИТМА

Рассмотрим масс-спектр $S$, в результате предварительной обработки которого получен массив спектральных пиков *P*. Параметры пиков заданы в соответствующих элементах массивов $\mu, I, \varepsilon$, где $\mu$ — положение центроида пика (отношение массы иона к заряду), $I$ — интенсивность пика, $\varepsilon$ — погрешность определения положений центроида пика. В результате обработки требуется определить совокупность параметров компонентов пробы: моноизотопных масс и интегральной интенсивности сигнала.

Специфика наложений изотопных мультиплетов в масс-спектрах пептидов заключается в преимущественном образовании неразрешенных дублетов, которые представляют собой парные наложения отдельных пиков, имеющих близкие отношения массы к заряду. Подобные дублеты обрабатываются как одиночные *составные* пики с интенсивностью, равной сумме интенсивности пиков, образующих наложение. Прочие спектральные пики, не подверженные наложению, именуются *несоставными*.

Предлагаемый алгоритм декомпозиции масс-спектра состоит из следующих операций:

1. Выбирается зарядовое состояние иона *z*.

2. Массив спектральных пиков *P* сканируется с целью формирования *тестовой группы* пиков, которая предположительно содержит изотопные мультиплеты компонентов пробы. Положения центроидов пиков тестовой группы пересчитываются для приведения к однозарядному иону и корректируются на величину $m_H$ молекулярной массы присоединенного носителя заряда: $m_t[d] = (\mu_t[d] - m_H)z$. При этом приближенно определяется молекулярная масса выделяемого компонента пробы.

3. В случае успешного формирования тестовой группы осуществляется ее интерпретация. Определяется модельный изотопный мультиплет, соответствующий молекулярной массе, рассчитанной на предыдущем шаге. Ключевыми операциями при интерпретации тестовой группы являются позиционирование и масштабирование модельного мультиплета. В результате выполнения этих операций определяются моноизотопный пик выделяемого изотопного мультиплета, моноизотопная масса и суммарная интенсивность сигнала выделяемого компонента пробы.

4. Для продолжения процедуры интерпретации интенсивности пиков выделенного мультиплета вычитаются из интенсивности пиков тестовой группы. В случае, если полученная разность обусловлена отличием фактической формы изотопного мультиплета от модельной и не превышает максимально допустимого отклонения, пик тестовой группы отмечается как интерпретированный и исключается из дальнейшей обработки. В противном случае пик требует дальнейшей интерпретации и возвращается в массив *P*, при этом интенсивность пика приравнивается полученной разности.

5. В случае, если после коррекции тестовая группа содержит пики с ненулевой интенсивностью, шаги 3 и 4 повторяются.

6. Производится переход к формированию следующей тестовой группы (шаг 2). В случае, если массив *P* не содержит более пиков, формирующих тестовые группы зарядового состояния *z*, производится переход к следующему зарядовому состоянию $(z-1)$ и шаги 2–6 повторяются.

7. По завершении анализа всех зарядовых состояний иона выделенные изотопные мультиплеты группируются по молекулярным массам. В результате образуется список компонентов пробы, для каждого из которых определены моноизотопная масса, суммарная интенсивность сигнала, а также перечень зарядовых состояний иона, обнаруженных в масс-спектре.

Рассмотрим подробнее процедуру интерпретации тестовой группы пиков (шаг 3). Позиционирование модельного изотопного мультиплета на тестовой группе пиков задается смещением *d* его моноизотопного пика относительно первого пика

тестовой группы и выполняется по-разному в зависимости от диапазона значений молекулярной массы. В молекулах пептидов невысокой молекулярной массы ($m$ < 3000 а.е.м.) моноизотопный пик является первым в порядке возрастания массы, обладает существенной относительной интенсивностью и уверенно детектируется в масс-спектре. В этом случае позиционирование модельного мультиплета осуществляется путем совмещения его первого пика с первым пиком тестовой группы, обладающим ненулевой интенсивностью. При этом смещение $d = 0$.

Масштабирование модельного изотопного мультиплета выполняется с целью приведения интенсивности его пиков в соответствие выделяемому изотопному мультиплету. Величина коэффициента масштабирования $\kappa$ определяется из условия совмещения несоставного пика тестовой группы с соответствующим *опорным* пиком модельного мультиплета:

$$\kappa = \min\left(\frac{I_t[k+d]}{I_p[k]}\right),$$

где $k = 1,...,n_\theta$; $I_t$, $I_p$ — интенсивности пиков тестовой группы и модельного изотопного мультиплета.

В случае, если в качестве опорного выбирается пик убывающего "шлейфа", характерного для изотопных мультиплетов пептидов, могут быть получены заниженные значения $\kappa$. Поэтому область поиска опорного пика ограничивается $n_\theta$ первыми пиками модельного изотопного мультиплета, суммарная интенсивность которых не превышает заданного значения $\theta$: $\sum_{i=1}^{n_\theta} I_p[i] < \theta$, $0 < \theta \le 1$.

По мере увеличения молекулярной массы пептида относительная интенсивность моноизотопного пика уменьшается, вместе с тем снижается вероятность его детектирования на фоне шума. Поэтому для пептидов с молекулярной массой $m$ > > 3000 а.е.м. позиционирование и масштабирование выполняются по условию минимума суммы квадратов отклонений интенсивности пиков тестовой группы и модельного мультиплета

$$\kappa = \frac{\sum_{k=1,...,n_\theta} I_t[k+d] I_p[k]}{\sum_{k=1,...,n_\theta} \left(I_p[k]\right)^2}.$$

Это позволяет, опираясь на форму изотопного мультиплета, восстановить положение моноизотопного пика, утерянного на предварительной стадии обработки масс-спектра.

Выполнение позиционирования и масштабирования модельного мультиплета по сути означает выделение изотопного мультиплета компонента пробы из тестовой группы пиков. Параметры выделенного компонента пробы определяются по результатам выполнения этих операций:

— моноизотопная масса молекулы компонента пробы: $m = m_t[d]$;

— интенсивности пиков изотопного мультиплета: $I[k] = I_p[k] \cdot \kappa$, где $k = 1,...,n_e$;

— суммарная интенсивность сигнала изотопного мультиплета: $I_\Sigma[z] = \kappa$.

Для продолжения процедуры интерпретации интенсивности пиков тестовой группы корректируются путем вычитания интенсивности пиков выделенного мультиплета. В случае, если полученная разность превышает максимально допустимое отклонение $I_t[k+d] - I_p[k] \cdot \kappa > \delta_{\max}[k]$, пик тестовой группы признается составным и возвращается в массив $P$ с остаточной интенсивностью $I = I_t[k+d] - I_p[k] \cdot \kappa$.

Величина $\delta_{\max}[k]$ рассчитывается на основании максимальной ошибки аппроксимации интенсивности спектральных пиков модельным изотопным мультиплетом ($\delta_{up}[k]$, $\delta_{dn}[h]$), а также значения $\rho$, ограничивающего динамический диапазон интенсивности изотопных мультиплетов, претерпевающих наложение:

$$\delta_{\max}[k] = \max\left(I_t[h+d]\left(\frac{I_p[k]+\delta_{up}[k]}{I_p[h]-\delta_{dn}[h]} - \frac{I_p[k]}{I_p[h]}\right);\ \max\left(I_p\right)\cdot\frac{\kappa}{\rho}\right),$$

где $h$ — индекс опорного пика.

Разработанный алгоритм декомпозиции был адаптирован для обработки массива данных (ВЭЖХ-МС)-эксперимента. При проведении такого эксперимента производится периодическая регистрация масс-спектров компонентов пробы, элюируемых из хроматографической колонки. В результате сигналы компонентов распределяются между несколькими последовательными масс-спектрами. Для обработки таких данных производится "скользящее" суммирование масс-спектров в границах времени, соответствующих времени выхода хроматографического пика. Массив полученных суммарных масс-спектров обрабатывается алгоритмом декомпозиции, и полученная информация о компонентах пробы заносится в массив предварительных результатов. Окончательный список компонентов пробы и их параметров определяется в результате решения задачи кластерного анализа на массиве предварительных результатов. Помимо молекулярной массы и суммарной интенсивности сигнала для каждого компонента пробы

рассчитывается осредненное время хроматографической элюции.

## РАСЧЕТ ПАРАМЕТРОВ МОДЕЛЬНОГО ИЗОТОПНОГО МУЛЬТИПЛЕТА ПЕПТИДОВ

Параметры модельного изотопного мультиплета рассчитывались на основе среднестатистических значений интенсивности пиков изотопных мультиплетов пептидов базы данных Swiss-Prot [11], в которой содержится наиболее достоверная информация о 163 496 белках и более чем 6 400 000 аминокислотных последовательностях пептидов (версия 45.1). Из множества пептидов БД Swiss-Prot были сформированы выборки, соответствующие ряду значений молекулярной массы. На каждой выборке были рассчитаны гистограммы плотности распределения относительной интенсивности пиков изотопных мультиплетов (рис. 2, а). Интенсивность пиков модельного изотопного мультиплета $I_p[k]$ рассчитывалась на основе выборочных значений математического ожидания (рис. 2, б). Для определения максимальной ошибки аппроксимации интенсивности пика ($\delta_{dn}[k], \delta_{up}[k]$) использовались квантили функций распределения порядка 0.01 и 0.99 (рис. 2, б).

В результате данного расчета получена таблица параметров модельного изотопного мультиплета для ряда значений молекулярной массы в диапазоне от 75 до 7000 а.е.м. Параметры модельного изотопного мультиплета в промежуточных значениях молекулярной массы рассчитываются методом линейной интерполяции.

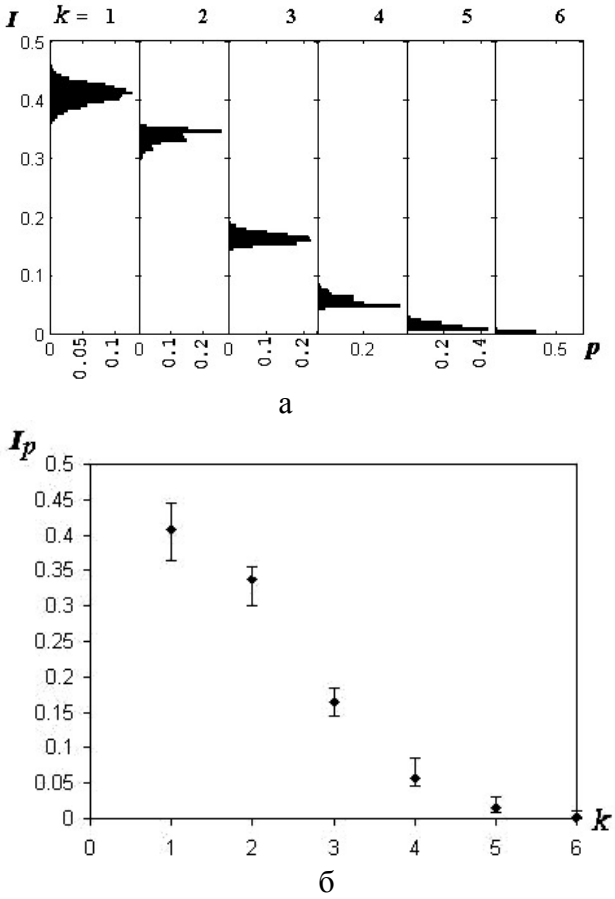


**Рис. 2.** Гистограммы плотности распределения относительной интенсивности пиков изотопных мультиплетов пептидов с молекуляной массой $m$=1500±10 а.е.м. (а) и параметры модельного изотопного мультиплета (б)

## ИССЛЕДОВАНИЕ ХАРАКТЕРИСТИК АЛГОРИТМА ДЕКОМПОЗИЦИИ

Алгоритм декомпозиции реализован в программном обеспечении обработки масс-спектрометрических данных в виде процедуры, написанной на языке C++. Для оценки качества работы алгоритма было произведено его тестирование на массиве модельных масс-спектрометрических данных. Качество декомпозиции масс-спектра характеризуется количеством корректно выделенных компонентов пробы, а также числом ложноположительных результатов. Ложноположительным результатом является выделение компонента, отсутствующего в пробе.

Для получения модельных масс-спектров была разработана программа, позволяющая генерировать масс-спектры проб заданного состава на основе аминокислотных последовательностей пептидов базы данных Swiss-Prot. Модельные масс-спектры генерировались в форме исходного сигнала с разрешением R = 10 000, в диапазоне отношения массы к заряду от 0 до 2500 а.е.м. и обрабатывались процедурой извлечения спектральных пиков. Это позволило учесть фактор ограниченной разрешающей способности прибора, воспроизвести наложения спектральных пиков и связанные с ними ошибки детектирования пиков, вносимые на предварительной стадии обработки. Массив выделенных спектральных пиков обрабатывался алгоритмом декомпозиции, и результат обработки сравнивался со списком компонентов пробы, на основе которого был генерирован модельный масс-спектр.

В результате обработки модельных масс-спектров были получены характеристики качества результатов выделения изотопных мультиплетов алгоритмом декомпозиции (рис. 3). Способность алгоритма к извлечению информации о компонентах, сигналы которых претерпевают наложения, иллюстрируют зависимости, представленные на рис. 4.

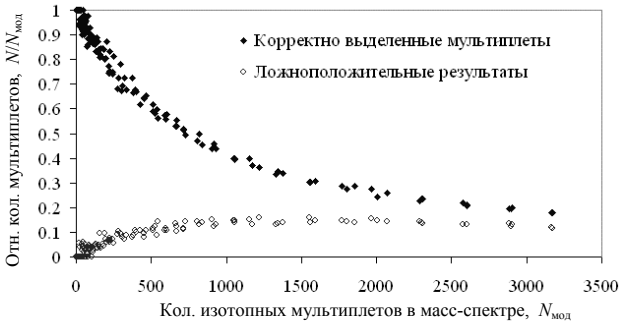


**Рис. 3.** Характеристики качества результатов выделения изотопных мультиплетов алгоритмом декомпозиции масс-спектра

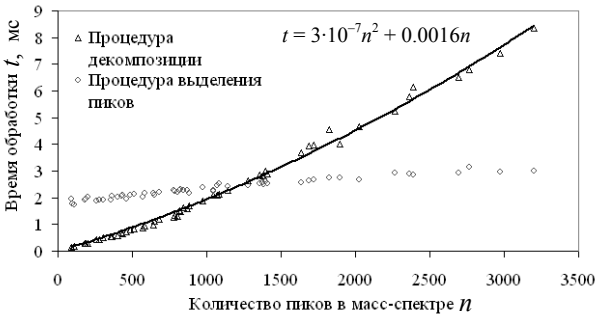


**Рис. 5.** Характеристика времени выполнения процедур обработки

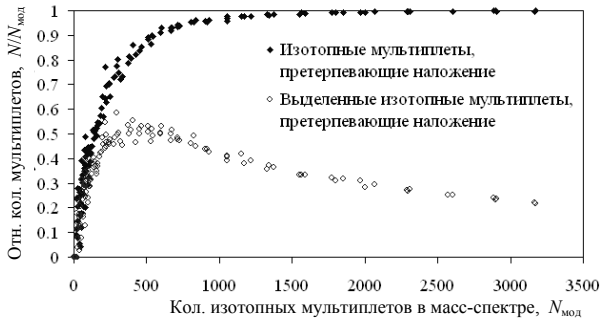


**Рис. 4.** Характеристика выделения изотопных мультиплетов, претерпевающих наложение

Для оценки вычислительной эффективности алгоритма была получена зависимость времени выполнения процедуры декомпозиции от количества пиков масс-спектра (рис. 5). Эксперимент проводился на компьютере с процессором Intel Pentium D с тактовой частотой 2800 МГц и объемом оперативной памяти 1500 МБ. Измерялось время выполнения процедур выделения пиков и декомпозиции модельных масс-спектров.

Полученная зависимость времени выполнения процедуры декомпозиции масс-спектра была аппроксимирована полиномом 2-й степени. Малая величина коэффициента при квадратичном члене (порядка $10^{-7}$) позволяет утверждать, что время выполнения алгоритма декомпозиции является практически линейным по отношению к числу обрабатываемых пиков. Слабая нелинейная составляющая объясняется использованием сортировки изотопных мультиплетов по молекулярным массам.

На заключительном этапе тестирования был произведен эксперимент по сравнению алгоритма декомпозиции с коммерческими аналогами, реализованными в современном программном обеспечении обработки масс-спектрометрических данных. Для проведения эксперимента был выбран пакет программ Analyst QC (продукт компании Applied Biosystems) с дополнительным модулем BioAnalyst, в котором реализованы следующие два алгоритма обработки масс-спектров с целью извлечения информации о компонентах пробы:

1. "Peak score reconstruct" ("Реконструкция на основе рейтинга пиков").
2. "Bayesian peptide reconstruct" ("Реконструкция на основе формул Байеса").

Эксперимент состоял в обработке модельных масс-спектров каждым из двух вышеперечисленных алгоритмов, а также алгоритмом декомпозиции и сравнении полученных результатов с составом модельных масс-спектров. С целью оценки производительности каждого из алгоритмов при обработке измерялось время, затраченное на вычисления.

На рис. 6 представлены диаграммы, иллюстрирующие относительное количество корректно выделенных компонентов, ложноположительных результатов, а также времени обработки модельных масс-спектров каждым из алгоритмов. Результаты эксперимента свидетельствуют, что алгоритм декомпозиции демонстрирует качество обработки, сопоставимое с коммерческими аналогами. При малом количестве компонент пробы алгоритм "Bayesian peptide reconstruct" обладает незначительным преимуществом, которое выражается в генерировании меньшего числа ложноположительных результатов. Однако с увеличением количества компонентов пробы проявляются преимущества алгоритма декомпозиции. Например, при количестве компонентов, равном 500, корректный результат наблюдается на 13 % чаще по сравнению с алгоритмом "Bayesian peptide reconstruct" и на 9 % чаще по сравнению с "Peak score reconstruct".

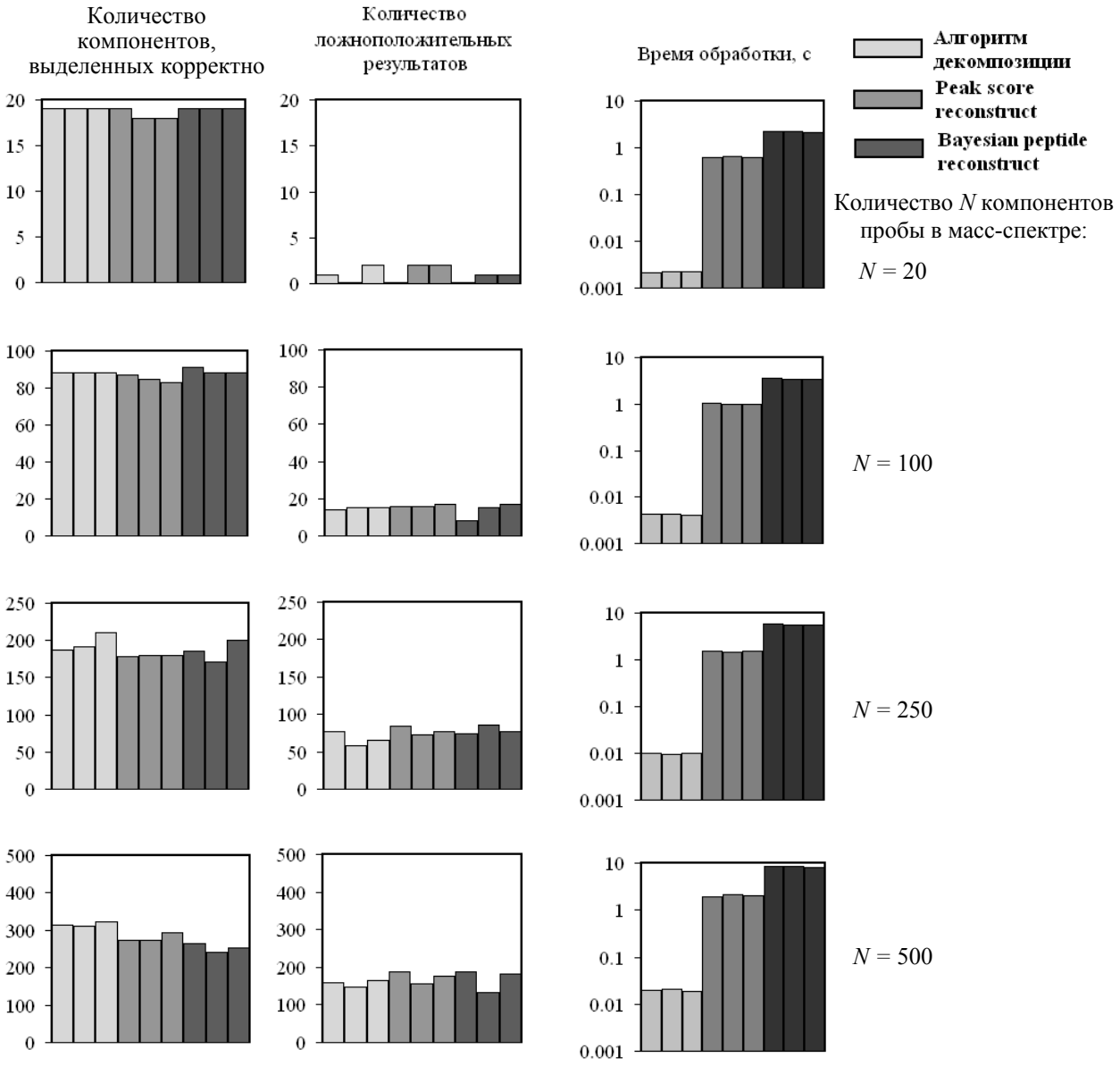


**Рис. 6.** Результаты тестирования алгоритмов обработки масс-спектров. Для каждого из алгоритмов показаны результаты обработки трех различных масс-спектров с одинаковым количеством компонентов пробы

При сопоставимом качестве обработки алгоритм декомпозиции демонстрирует значительно более высокую производительность. Общее время обработки модельных масс-спектров алгоритмами выделения пиков и декомпозиции составляло от 5 до 19 мс, что в 10–100 раз меньше, чем при обработке алгоритмами пакета BioAnalyst.

Столь заметное преимущество в производительности достигнуто за счет выделения информации о структуре масс-спектрометрического сигнала, а также за счет применения процедуры выделения изотопных мультиплетов, основанной на использовании модельного мультиплета.

## ОБРАБОТКА МАССИВА ДАННЫХ (ВЭЖХ-МС)-ЭКСПЕРИМЕНТА

Для тестирования алгоритма декомпозиции был поставлен эксперимент по идентификации белка методом пептидного массового картирования ("peptide mass fingerprint", PMF). В качестве модельного белка был выбран бычий сывороточный

альбумин (БСА), широко используемый для тестирования аналитических характеристик масс-спектрометрического оборудования.

Препарат БСА (Sigma, США) был гидролизован трипсином, затем полученная смесь пептидов анализировалась методом ВЭЖХ-МС. Масс-спектрометрический анализ проводился на времяпролетном масс-спектрометре МХ-5303 с источником ионов "электроспрей" (разработка Института аналитического приборостроения РАН). Предварительное разделение пробы производилось на жидкостном хроматографе "Милихром А-02" (производства ЗАО Институт хроматографии "Эконова", г. Новосибирск), работающем в режиме прямой стыковки с источником ионов.

В результате масс-спектрометрического эксперимента был получен массив из 569 масс-спектров, регистрация которых производилась в течение 21 мин с периодом около 2 с.

Массив экспериментальных данных был обработан алгоритмом декомпозиции, адаптированным для обработки (ВЭЖХ-МС)-данных, за время, равное 6.21 с. Список из 367 компонентов пробы, полученный в результате обработки, был направлен в программу Mascot, доступную через web-интерфейс по адресу http://www.matrix-science.com для интерпретации по методу пептидного массового картирования. Проба была верно интерпретирована как триптический гидролизат БСА. Данный вариант интерпретации оценен наивысшим значением рейтинга, в то время как оценки прочих гипотез лежат ниже порога достоверности, вычисляемого программой Mascot.

Пептиды, обнаруженные в пробе, в совокупности покрывают 71 % аминокислотной последовательности белка, что является высоким показателем для метода пептидного массового картирования. Высокая достоверность идентификации белка свидетельствует о качественном решении задачи извлечения аналитически значимой информации из масс-спектрометрических данных

## ЗАКЛЮЧЕНИЕ

Представленный алгоритм декомпозиции масс-спектра может применяться для обработки масс-спектров многокомпонентных проб экспериментов протеомики. При сопоставимом качестве обработки алгоритм превосходит коммерческие аналоги по производительности более чем в 10 раз и демонстрирует зависимость времени обработки от числа спектральных пиков, характер которой близок к линейному. Благодаря использованию модельного изотопного мультиплета пептида алгоритм позволяет восстановить спектральные пики, которые не были выделены на предварительной стадии обработки.

Разработанный алгоритм реализован в программном обеспечении системы регистрации и обработки результатов эксперимента времяпролетного масс-спектрометра МХ-5303, разработанного в Лаборатории экологической и биомедицинской масс-спектрометрии Института аналитического приборостроения РАН по госконтракту ОКР № 40.032.11.17.

**Благодарности**

Авторы выражают глубокую благодарность А. Подтележникову, А. Новикову, И. Краснову и Е. Подольской, предоставившим масс-спектрометрические данные (ВЭЖХ-МС)-экспериментов.

## СПИСОК ЛИТЕРАТУРЫ

1. *Tyers M., Mann M.* From genomics to proteomics // Nature. 2003. V. 422. P. 193–197.
2. *Aebersold R., Mann M.* Mass spectrometry-based proteomics // Nature. 2003. V. 422. P. 198–203.
3. *Mann M., Meng C.K., Fenn J.B.* Interpreting mass spectra of multiply charged ions // Analytical Chemistry. 1989. V. 61. P. 1702–1708.
4. *Fenn J.B., Mann M., Meng C.K.* Патент США № 5 130 538, 1992.
5. *Hagen J.J., Monning C.A.* Method of estimating molecular mass from electrospray mass spectra // Analytical Chemistry. 1994. V. 66. P. 1877–1883.
6. *Reinhold B.R., Reinhold V.N.* Electrospray ionisation mass spectrometry: deconvolution by an entropy-based algorithm // J. Am. Soc. Mass Spec. 1992. V. 3. P. 207–215.
7. *Ferrige A.G. et al.* Maximum entropy deconvolution in electrospray mass spectrometry // Rapid Comm. Mass Spec. 1991. V. 5. P. 374–377.
8. *Ferrige A.G. et al.* Disentangling electrospray spectra with maximum entropy // Rapid Comm. Mass Spec. 1992. V. 6. P. 707–711.
9. *Zhang Z., Marshall A.G.* A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra // J. Am. Soc. Mass Spec. 1998. V. 9. P. 225–233.
10. *Pearcy J.O., Lee T.* MoWeD, a computer program to rapidly deconvolute low resolution electrospray liquid chromatography/mass spectrometry runs to determine component molecular weights // J. Am. Soc. Mass Spec. 2001. V. 12. P. 599–606.
11. *Bairoch A., Apweiler R.* The SWISS-PROT proteine sequence database and its supplement TrEMBL in 2000 // Nucleic acid research. 2000. V. 28. P. 45–48.

***Институт аналитического приборостроения РАН, Санкт-Петербург*** (*Макаров В.В., Лютвинский Я.И., Веренчиков А.Н., Краснов Н.В.*)

***Балтийский государственный технический университет, Санкт-Петербург*** (*Савельев С.К.*)

Материал поступил в редакцию 17.04.2006.

# DATA MINING ALGORITHM FOR MASS SPECTRA OF PROTEOMIC EXPERIMENT

**V. V. Makarov, S. K. Saveliev*, Ya. I. Lutvinsky, A. N. Verenchikov, N. V. Krasnov**

*Institute for Analytical Instrumentation RAS, Saint-Petersburg*
**Baltic State Technical University, Saint-Petersburg*

A data mining algorithm is described for extracting analytically significant information from mass spectra of complex peptide mixtures. The algorithm operates with mass-to-charge and intensity values of mass spectral peaks, and can correctly process overlapped isotope clusters. The algorithm is shown to be much faster than commercial ones, while the quality of processing is shown to be the same.