На правах рукописи

ЛЮТВИНСКИЙ

Ярослав Игоревич

МЕТОД РАСПОЗНАВАНИЯ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В МАСС-СПЕКТРАХ ПЕПТИДОВ ДЛЯ ЗАДАЧ ПРОТЕОМИКИ

СПЕЦИАЛЬНОСТЬ: 01.04.01 – ПРИБОРЫ И МЕТОДЫ ЭКСПЕРИМЕНТАЛЬНОЙ ФИЗИКИ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

САНКТ-ПЕТЕРБУРГ

2007

Работа выполнена в Институте аналитического приборостроения Российской академии наук.

Научный руководитель: кандидат технических наук
Новиков Лев Васильевич

Официальные оппоненты: доктор химических наук, профессор
Зенкевич Игорь Георгиевич (СПбГУ)

кандидат физико-математических наук
Бердников Александр Сергеевич (ИАнП РАН)

Ведущая организация: Филиал Института энергетических проблем химической физики Российской академии наук

Защита состоится "_27_" декабря 2007г. в _$15^{\underline{00}}$_ часов на заседании Диссертационного Совета Д002.034.01 при Институте аналитического приборостроения РАН по адресу: 190103, Санкт-Петербург, Рижский пр., 26.

С диссертацией можно ознакомиться в научно-технической библиотеке ИАнП РАН по адресу: 190103, Санкт-Петербург, Рижский пр., 26.

Автореферат разослан "25" ноября 2007 г.

Ученый секретарь
диссертационного совета Д002.034.01, А.П.Щербаков
кандидат физико-математических наук

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Одной из наиболее динамично развивающихся областей современной молекулярной биологии является протеомика – исследование белкового пула организма (протеома) как единого целого. К числу ведущих методологий в протеомных исследованиях относится масс-спектрометрия высокого разрешения с мягкими методами ионизации. На нужды протеомики ориентирована, в значительной степени, разработка новых современных тандемных масс-спектрометров. Появление новых приборов вызывает необходимость в разработке новых методов обработки информации, получаемой на этих приборах.

Как правило, тандемные масс-спектрометры в протеомике используются для анализа смесей белков, представленных продуктами избирательного ферментативного гидролиза. Получаемые масс-спектры представляют собой фрагментные масс-спектры пептидов – продуктов гидролиза. Важнейшая задача при обработке получаемых данных - это восстановление аминокислотной последовательности пептида по его фрагментному спектру.

Одним из перспективных, но пока недостаточно алгоритмически проработанных подходов к интерпретации фрагментных масс-спектров является частичное восстановление аминокислотной последовательности по наблюдаемым в спектрах сериям основных фрагментов пептида. Такая методика интерпретации фрагментных масс-спектров получила в мировой научной литературе название Peptide Sequence Tag (PST) Search. Этот подход к интерпретации масс-спектров имеет следующие достоинства:

- высокая скорость интерпретации;
- устойчивость результата интерпретации масс-спектра по отношению к пост-трансляционным модификациям, точечным мутациям, неполному и неспецифичному гидролизу;
- высокая надежность получаемых результатов интерпретации, обусловленная использованием информации, действительно присутствующей в масс-спектре.

Преимущества стратегий обработки данных, основанных на методе поиска PST, обеспечили широкое распространение этого метода интерпретации масс-спектров среди биологов. Эти стратегии обеспечивают:

- более полное использование масс-спектрометрической информации за счет распознавания спектров модифицированных пептидов
- идентификацию белков на основе спектров низкого качества, содержащих малое количество информативных сигналов и большое количество шума
- идентификацию пост-трансляционных модификаций белка
- идентификацию ближайших гомологов исследуемого белка

К сожалению, до последнего времени не существовало удачных реализаций алгоритмов поиска PST и, часто, распознавание PST проходит вручную, порождая большое количество монотонной работы.

Только в самое последнее время появились алгоритмы, удачно автоматизирующие частичное восстановление аминокислотной последовательности пептидов. Однако, одним из существенных недостатков существующих алгоритмов является то, что каждый алгоритм разрабатывается для конкретного класса приборов, и не может быть впоследствии адаптирован к приборам другого класса.

Целью работы является разработка высокоэффективного адаптивного метода распознавания аминокислотной последовательности пептида во фрагментном масс-спектре.

Для достижения этой цели предложен высокопроизводительный алгоритм распознавания аминокислотной последовательности пептида во фрагментном масс-спектре и предложена процедура оценки критериев значимости спектральной информации в фрагментных масс-спектрах.

Научная новизна работы

1. Предложена методика численной оценки значимости эмпирических критериев для использования масс-спектрометрической информации при решении задачи распознавания аминокислотной последовательности пептида в его фрагментном масс-спектре.
2. Предложен и апробирован метод ранжирования гипотез об аминокислотной последовательности пептида, построенных на основании фрагментного масс-спектра.
3. Предложен новый алгоритм распознавания аминокислотной последовательности пептида во фрагментном масс-спектре, оптимизированный по числу проверяемых гипотез.

Практическая значимость работы

Разработан высокопроизводительный адаптивный алгоритм распознавания аминокислотной последовательности пептида во фрагментном масс-спектре, названный CrystalTag. Этот алгоритм может использоваться для обработки массивов фрагментных масс-спектров пептидов в экспериментах протеомики, проведенных на масс-спектрометрических приборах различных типов.

Предложенный алгоритм обладает следующими достоинствами:

— **Быстродействие**. Благодаря оптимизированному по числу проверяемых гипотез способу анализа масс-спектра, время обработки спектра алгоритмом CrystalTag составляет менее миллисекунды, что намного меньше времени регистрации спектра на существующих тандемных масс-спектрометрах.

— **Качество распознавания**. Алгоритм дает высокую вероятность наличия достоверной гипотезы среди предложенных гипотез.

— **Адаптивность**. Предложенная процедура оценки модели фрагментации позволяет использовать алгоритм для масс-спектров, полученных на масс-спектрометрах различной конструкции, использующих разные физические принципы и имеющих различные аналитические характеристики.

— **Расширяемость**. Байесова модель формирования оценки гипотез позволяет легко вводить новые критерии, значимые для восстановления исходной последовательности пептидов.

Алгоритм реализован в составе программного комплекса автоматической обработки данных фрагментных масс-спектров, полученных в экспериментах протеомики. Программный комплекс предназначен для получения биологически значимого ответа на основании массива фрагментных масс-спектров.

Положения, выносимые на защиту

1. Метод численной оценки значимости эмпирических критериев для использования масс-спектрометрической информации при решении задачи распознавания аминокислотной последовательности пептида во фрагментном масс-спектре.
2. Метод ранжирования гипотез об аминокислотной последовательности пептидов, распознанных во фрагментном масс-спектре этих пептидов.
3. Алгоритм построения гипотез об аминокислотной последовательности пептидов по фрагментному масс-спектру пептида.

Апробация работы. Результаты работы были доложены на конференции «Аналитическое приборостроение» (Санкт-Петербург, 2005г.), на II съезде Всероссийского

масс-спектрометрического общества (Москва, 2005г.), на III съезде Общества биотехнологов России (Москва, 2005г.), на международной выставке «Biotechnica 2005» (Ганновер, 2005г.).

Структура и объем диссертации. Диссертация состоит из введения, обзора литературы, постановки задачи на разработку методов и алгоритма, изложения разработанных методов и алгоритма, описания программного комплекса, содержащего реализацию методов и алгоритма, описания и обсуждения результатов его тестирования, заключения и списка используемых источников. Диссертация изложена на130 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Природа данных фрагментных масс-спектров пептидов

Основным способом получения фрагментных масс-спектров пептидов является применение метода тандемного масс-спектрометрического анализа. Тандемная масс-спектрометрия (МС-МС) используется для структурного анализа и идентификации веществ в составе смесей.

Методика МС-МС состоит из следующих операций.

- Разделение в первой МС-ступени первичных, или "родительских", ионов и селекция ионов с единственным значением отношения массы к заряду (m/z).
- Фрагментация родительских ионов с образованием разнообразных структурно значимых ионных фрагментов, называемых вторичными, или "дочерними", ионами.
- Масс-анализ дочерних ионов.

Существенно, что в случае применения тандемной масс-спектрометрии к смесям пептидов каждый отдельный масс-спектр отображает набор фрагментов ***одного пептида***.

Тандемные масс-спектрометры используют различные физические принципы для разделения родительских ионов, фрагментации и масс-анализа дочерних ионов.

Состав, интенсивность сигналов, точность определения масс фрагментов, отражающих структуру анализируемого вещества, напрямую зависит от используемого прибора. В настоящее время для фрагментации наиболее часто используется явление столкновительной диссоциации (Collisionaly Induced Dissociation - CID). В ячейке CID ионы сталкиваются с нейтральными молекулами газа, заполняющего ячейку, что приводит к разрыву ковалентных связей в полипептидной цепи.

Для фрагментных ионов пептидов принята классификация, предложенная в 1984 году в работе [1], в соответствии с которой фрагменты, содержащие N-концевую аминокислоту, в зависимости от разорванной ковалентной связи обозначают латинскими буквами a, b и c. Подобно, фрагменты С-конца обозначены x, y и z. Каждый обозначенный ион снабжается индексом,– индекс соответствует числу аминокислотных остатков оставшихся при ионе после фрагментации.

Кроме разрыва полипептидной цепи при фрагментации происходят и такие процессы как фрагментация по боковым радикалам, многократная фрагментация и множество других процессов, поэтому масс-спектр содержит множество сигналов кроме основных предполагаемых. Тем не менее, в спектре, полученном на хорошо настроенном масс-спектрометре, как правило, доминируют b и y ионы.

Глава 2 Методы биохимической интерпретации фрагментных масс-спектров пептидов

Восстановление аминокислотной последовательности пептида по его фрагментному масс-спектру – это один из основных приемов интерпретации масс-спектрометрических данных, используемых в протеомике.

В настоящее время наиболее распространенным подходом к задаче восстановления аминокислотной последовательности является поиск наилучшего совпадения состава

сигналов в зарегистрированных экспериментальных фрагментных спектрах и теоретических спектрах, построенных на основании аминокислотных последовательностей известных белков, содержащихся в протеомных базах данных. Этот метод картирования пептидных фрагментов реализован в таких известных программных комплексах как Mascot, Sequest, X!Tandem и другие.

Альтернативный подход подразумевает восстановление аминокислотной последовательности без обращения к базам данных на основании непосредственного анализа сигналов спектров. В рамках этой альтернативы можно выделить два метода:

- Полное восстановление аминокислотной последовательности – подход, реализованный в таких программных продуктах как Lutefisk, Peaks Studio. Отсутствие во многих фрагментных масс-спектрах даже хорошего качества полных серий сигналов, соответствующих сериям фрагментных ионов, приводит к базовому недостатку этого метода: восстановленная последовательность содержит слабые предположения, основанные на неполной информации и, поэтому, часто не соответствует действительной аминокислотной последовательности.

- Частичное восстановление аминокислотной последовательности. Для частичного восстановления используются наблюдаемые в спектре последовательности пиков, принадлежащих основным сериям фрагментов пептида, расстояние между которыми соответствует массам аминокислотных остатков пептида. Как правило, такая последовательность пиков покрывает не весь спектр и позволяет восстановить только часть аминокислотной последовательности пептида. Данная работа посвящена выявлению наиболее вероятных последовательностей пиков, представляющих аминокислотную последовательность исходного пептида.

Идея использовать эти последовательности пиков для поиска в базах данных белков впервые высказана в статье [2] и техника такого поиска в мировой литературе получила название Peptide Sequence Tag Search.

Таким образом, Peptide Sequence Tag (PST) – это последовательность пиков во фрагментном масс-спектре пептида, трактуемая как отображение части аминокислотной последовательности пептида на фрагментный масс-спектр за счет отнесения пиков последовательно к одной серии фрагментных ионов пептида.

PST принято записывать как массу первого пика в последовательности, последовательность аминокислот, соответствующую расстояниям между пиками последовательности, и разницу между последним пиком последовательности и массой родительского иона, например PST на рисунке 1 можно обозначить как *[611.30]LGADE[242.05]*.

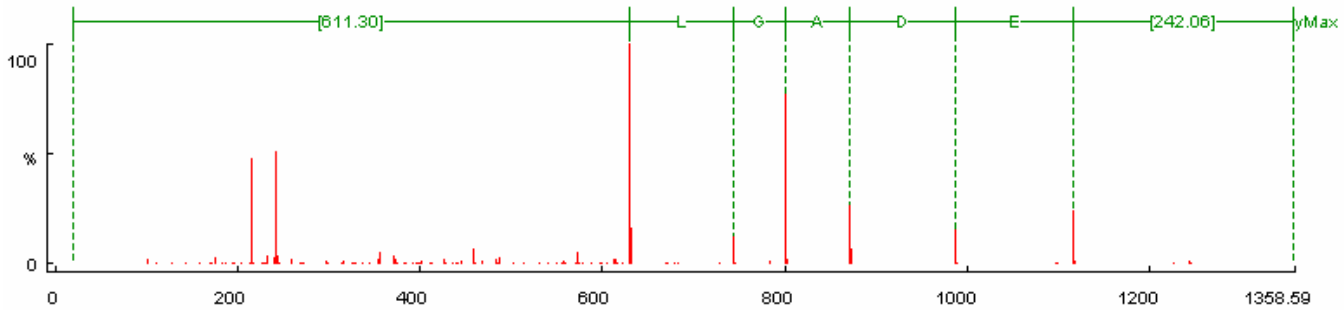

Рис. 1*. Спектр пептида IEEDAGLGNGGLLGR с выделенным PST [611.30]LGADE[242.05]*

Необходимо отметить, что по последовательности пиков, в общем случае не удается установить направление последовательности – от N-конца к C-концу или наоборот, поскольку неизвестно, ионы какой именно серии фрагментов (b или y) представлены пиками. Также на основании тэга не удается различить аминокислоты лейцин и изолейцин с одинаковой молекулярной массой, и часто (при недостаточной точности определения масс)

не удается различить аминокислоты глютамин и лизин с близкими молекулярными массами (разница 0.03 Да).

В настоящее время ведутся интенсивные работы как в области разработки новых алгоритмов построения PST, так и использования PST для идентификации и характеризации белков в ходе белковых анализов.

Интерес к алгоритмам построения и использования PST обусловлен следующими причинами:

- PST, выявленные в результате анализа фрагментных масс-спектров пептидов, полученных в результате неполного или неспецифического гидролиза, а также содержащих пост-трансляционные модификации, тем не менее, позволяют использовать спектр для идентификации белков.
- PST пригодны для идентификации функций неизвестных белков у организмов с несеквенированным геномом на основании установления ближайших исследованных гомологов белка
- PST обладают меньшей информационной избыточностью по сравнению со списком пиков масс-спектра, что снижает время поиска в протеомных базах данных и позволяет использовать PST для создания систем быстрой обработки масс-спектрометрических данных белковых анализов.

Глава 3. Разработка алгоритма частичного восстановления последовательности пептида по его фрагментному масс-спектру

Постановка задачи: Алгоритм распознавания PST формализуется как поиск частичного пути во взвешенном ориентированном ациклическом графе [3]. В таком графе вершины представлены сигналами масс-спектра, а ребра допустимых переходов – разницами масс, соответствующих массам аминокислотных остатков (см Рис. 2). На рисунке сплошной черной линией выделен корректный путь через граф, пунктирными линиями несколько вариантов ложно-позитивных результатов.

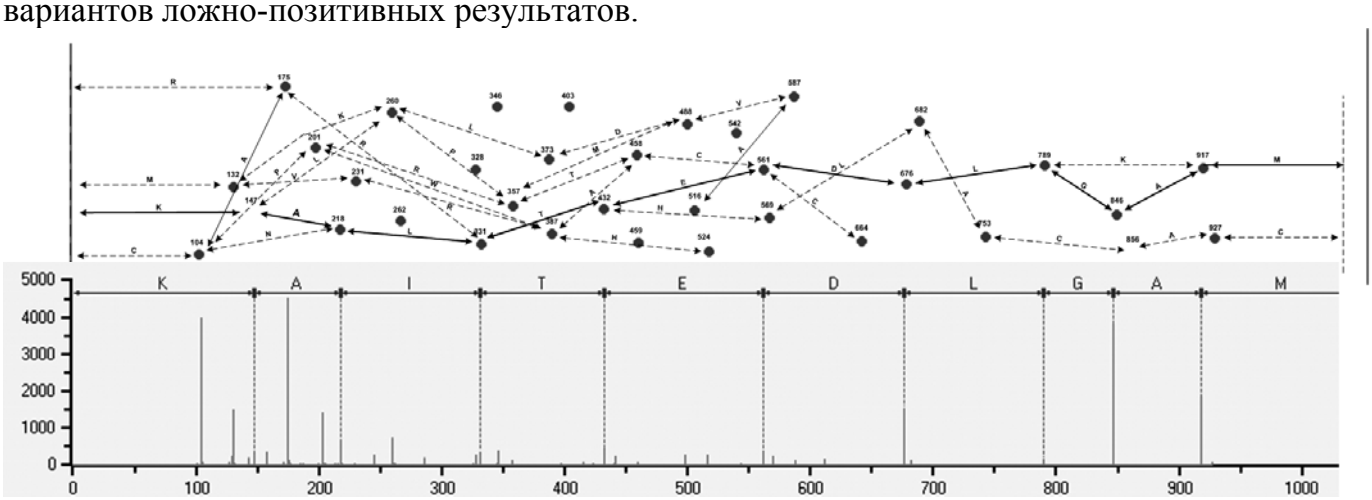

Рис. 2 *Представление масс-спектра как графа на примере масс-спектра пептида MAGLDETIAK (с упрощением)*

Классические алгоритмы поиска наилучшего пути в графе, такие как алгоритм Дийкстры или алгоритм A-Star, оперируют понятием фиксированной начальной и конечной точек пути. В нашем случае, хотя начальные и конечные точки детерминированы (нулевая отметка массовой шкалы и масса родительского иона), в их достижении нет необходимости. Более того, правильного решения задачи нахождения полного пути для такого графа может не существовать – немногие спектры содержат полные серии фрагментных ионов. Из-за этого подход классических алгоритмов поиска пути малоприменим.

В современных алгоритмах поиска PST для построения списка проверяемых гипотез используются различные варианты поиска в глубину, применяемые относительно каждого

узла графа. Это приводит к полному перебору гипотез, число которых достигает десятков тысяч на спектр.

Таким образом, при решении задачи частичного восстановления аминокислотной последовательности пептида можно выделить две основные исследовательские задачи:

1. Построение адекватных оценок для гипотез, обнаруживаемых при анализе графа, должно позволить из множества гипотез выделить верную гипотезу.
2. Алгоритм построения оптимального пути должен быть оптимизирован по отношению к числу проверяемых гипотез, поскольку проверка всех существующих гипотез приводит к потере времени выполнения алгоритма на проверку заведомо ложных гипотез.

Оценка графа масс-спектра:

Для выбора критериев оценки пиков воспользуемся опытом экспертов, проводящих распознавание PST вручную. Для оценки возможности вхождения пика в PST есть ряд эмпирических критериев, таких как:

- Относительная интенсивность пика в его окрестности
- Зашумленность спектра вокруг пика в его окрестности
- Наличие в спектре пиков, парных данному, по правилам построения серий ионов – y↔b, y↔a, x↔b и т. д.
- Наличие характерных нейтральных потерь $-H_2O$, $-NH_2$, и т. д.

Ни один из этих признаков не является определяющим. Для того, чтобы оценить правдоподобие гипотезы о принадлежности пика одной из основных серий ионов, требуется комплексная многокритериальная оценка. Для выполнения этой задачи, а также и для оценки степени значимости перечисленных критериев воспользуемся многокритериальной оценкой на основе теоремы Байеса.

Для построения этой оценки этого нам потребуется оценить условные вероятности $P(A_i|H)$ выполнения каждого из критериев A_i при условии выполнения двух гипотез, образующих полный набор:

– H_1 – пик относится к серии фрагментных ионов b или y.

– H_2 – пик не относится к серии фрагментных ионов b или y.

Решение о принадлежности или не принадлежности сигнала к серии фрагментных ионов принимается, исходя из восстановления теоретической картины фрагментации для масс-спектров известных пептидов.

Под выполнением критерия будем подразумевать:

- наличие пика в соответствующей позиции масс-спектра, для таких критериев как наличие парных пиков и нейтральных потерь. Наличие или отсутствие пика образует полный набор событий, возможных при анализе данных критериев.

- для критериев относительной интенсивности и зашумленности спектра, значение которых оценивается числом, выберем набор интервалов, также покрывающих полный набор событий, после чего вычислим условную вероятность для гипотез H_1, H_2 для каждого интервала.

Собранный набор условных вероятностей $P(A_i|H)$ позволяет оценить по Байесу вероятность гипотезы H_1 для каждого пика спектра, в том случае, если мы можем предполагать, что значимость критериев для этого спектра адекватна доверяемым данным, использованным для накопления статистики. Оценку каждого пика мы получаем последовательным применением формулы Байеса для каждого из предварительно оцененных критериев A_i.

$$P(H_1 \mid A_i) = \frac{P(H_1)P(A_i \mid H_1)}{P(H_1)P(A_i \mid H_1) + (1 - P(H_1))P(A_i \mid H_2)} \qquad (1)$$

При первом применении теоремы Байеса в качестве априорной вероятности $P(H_1)$ используется доля сигналов ионов серий y и b в спектрах пептидов. При последовательной оценке по ряду критериев в качестве априорной вероятности используется апостериорная вероятность, полученная на предыдущем шаге.

В качестве итоговой оценки вершины графа масс-спектра Q_j используется апостериорная вероятность, полученная после применения всех оцененных критериев.

Оценка ребер графа, построенного на основании масс-спектра, сводится к оценке допустимости предположения о том, что разность масс между двумя пиками является измерением массы аминокислотного остатка. Для оценки допустимости этого предположения используется нормальное распределение ошибки измерения разницы масс между фрагментными ионами y и b серий.

$$p(\delta) = \exp(-\delta^2 / 2k\sigma^2) \qquad (2)$$

Где δ – наблюдаемая погрешность точной массы аминокислотного остатка для интервала между пиками; σ – численно оцениваемое среднеквадратичное отклонение измерения точной массы аминокислотного остатка для пиков, отнесенных к y- и b- сериям ионов на основании доверяемых данных.

Поскольку оценка $p(\delta)$ имеет не байесовый характер, для уравновешивания влияния этой оценки на результаты работы алгоритма был введен параметр k. По результатам исследований выяснилось, что результаты работы алгоритма в значительной степени устойчивы к изменению параметра k в интервале от 1 до 10. Рекомендованное значение параметра: k=3.

Оценку Peptide Sequence Tag (PST) в целом будем строить как произведение оценок всех вершин и ребер графа масс-спектра, вошедших в PST. Таким образом, итоговая оценка построенного PST из n пиков будет:

$$P_{tag} = \left(\prod_{j=1}^{n} Q_j \right) \left(\prod_{j=1}^{n-1} p(\delta_j^{j+1}) \right) \qquad (3)$$

В терминах теории вероятностей эта оценка соответствует совпадению событий включения в путь PST вершин и ребер графа. На взгляд автора, это соответствует нарастанию вероятности ошибки при увеличении длины PST.

В главе 5 показано, что полученное значение является адекватной оценкой гипотезы, так как, чем меньше оценка, тем меньше фактическая достоверность гипотезы.

Статистическое исследование масс-спектров

Для построения набора условных вероятностей необходимо использовать масс-спектры известных пептидов. В качестве таких ***доверяемых данных*** будем использовать результаты интерпретации таких систем как X!Tandem и Mascot.

Поскольку одной из целей разработки является универсальность алгоритма, для исследования были привлечены выборки данных, полученные на масс-спектрометрах, имеющих существенно разные аналитические параметры и построенных на основании различных физических принципов:

1-я выборка масс-спектров получена из репозитория масс-спектрометрических данных Института Системной Биологии (Institute for Systems Biology, Seattle, USA) http://sashimi.sourceforge.net/repository.html. Выборка была получена в результате ВЭЖХ-МС-МС анализа модельной смеси 18 известных белков на приборе Q-TOF Ultima (Waters, США).

2-я выборка данных получена в результате ряда ВЭЖХ-ВЭЖХ-МС-МС экспериментов на масс-спектрометре Bruker Esquire (Ion Trap MS) в процессе белковых анализов препарата митохондрий клеток сердца быка, проведенных в Институте Биоорганической Химии РАН.

3-я выборка масс-спектров составлена в университете Упсалы (Швеция) в группе Биомедицинской масс-спектрометрии под руководством проф. Р. А. Зубарева. Выборка составлена по результатам ряда ВЭЖХ-МС/МС анализов, проведенных на приборе LTQ-FT для различных препаратов белков E. Coli и H. Sapiens.

Таблица 3.1 демонстрирует несколько разный набор значимых критериев для приборов различной архитектуры. Некоторые критерии дают высокую избирательность для всех приборов. Так критерий образования парных пиков y↔b свидетельствует о том, что вероятность образования парных пиков y↔b на порядок выше для ионов основных серий, нежели чем для случайных ионов. В то же время, критерии нейтральных потерь $-H_2O$, $-NH_3$ более значимы для времяпролетного прибора Q-TOF Ultima, нежели чем для приборов, в которых фрагментация происходит в ионной ловушке – Bruker Esquire и Finnigan LTQ-FT. Некоторые критерии не подтвердили своей информативности. Например, наличие ионов x-серии не может служить для оценки пиков, так как вероятность обнаружения таких ионов не коррелирует с природой оцениваемого пика.

Таблица 1. *Условные вероятности реализации критериев для гипотез H_1 и H_2*

	Q-TOF Ultima (Выборка №1)		Bruker Esquire (Выборка №2)		Finnigan LTQ-FT (Выборка №3)	
Критерии A_i:	$P(A_i \mid H_1)$	$P(A_i \mid H_2)$	$P(A_i \mid H_1)$	$P(A_i \mid H_2)$	$P(A_i \mid H_1)$	$P(A_i \mid H_2)$
y↔b	0.4938	0.03545	0.326	0.03671	0.63405	0.04207
y↔a	0.1573	0.05489	0.04309	0.03507	0.06871	0.05799
b↔x	0.04335	0.03174	0.026	0.03117	0.02413	0.03513
b↔a	0.1965	0.2819	0.04045	0.03982	0.07458	0.04666
y↔x	0.08348	0.2819	0.01934	0.03982	0.02627	0.04666
$-H_2O$	0.3514	0.247	0.1865	0.0511	0.3374	0.1017
$-NH_3$	0.319	0.202	0.0775	0.03586	0.1888	0.07728

Рисунок 3 показывает распределение вероятностей для критериев, заданных на наборах интервалов значений оцениваемого критерия. Для всех типов приборов эти два критерия оказываются существенно значимыми, то есть интенсивные пики и пики, расположенные в незашумленных областях спектра, будут предпочтительны для построения PST.

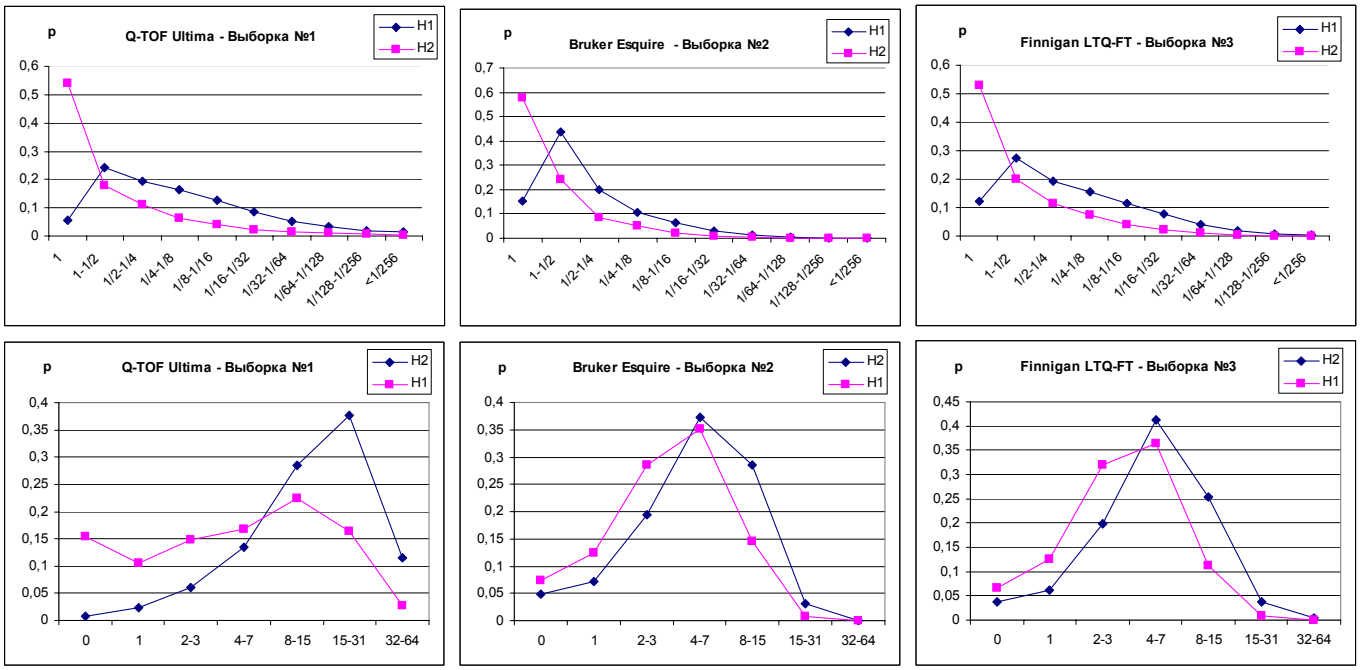

Рис. 3. *Относительная интенсивность (верхние диаграммы) и зашумленность (нижние диаграммы) спектра для информативных и неинформативных ионов*

Построение PST

Идея алгоритма CrytsalTag – использовать первыми лучшие данные масс-спектра. Для этого необходимо получить сортированный список однородных структур данных, соответствующих элементарным гипотезам о включении пика в PST, что подразумевает:

a) пик относится к одной из основных серий ионов b или y – оценка предположения Q_j
b) в спектре есть пик соответствующий следующему иону той же серии, в направлении возрастания масс – оценка предположения $p(\delta_j^{j+1})$
c) в спектре есть пик соответствующий следующему иону той же серии, в направлении убывания масс оценка предположения $p(\delta_{j-1}^{j})$

Если рассматривается только одно из двух последних условий, то данный пик – это конечный пик в PST.

Построим для каждого пика спектра полный набор структур данных, соответствующих всем элементарным гипотезам о включении пика в PST, в том числе и для завершения PST этим пиком. Оценим каждую из этих гипотез как произведение оценок пика и квадратных корней оценок интервалов.

$$G_j = \sqrt{p(\delta_{j-1}^{j})}Q_j\sqrt{p(\delta_j^{j+1})} \qquad (4)$$

Упорядочим получившийся набор структур по убыванию оценок. Заметим, что любой PST может быть представлен как цепочка таких структур, замкнутая со стороны убывания масс гипотезой, для которой не рассматривается предположение (c), а со стороны возрастания масс гипотезой, для которой не рассматривается предположение (b). Оценка PST будет равна произведению оценок структур его составляющих.

Далее извлекаем структуры из упорядоченного списка, и для каждой структуры строим все варианты цепочек структур с участием предыдущих извлеченных структур. Те цепочки, длина которых соответствует заранее заданному требуемому числу аминокислот в PST, и которые завершены по концам односторонними структурами, рассматриваем как итоговые версии PST для данного спектра.

Благодаря монотонному убыванию оценок рассматриваемых гипотез в каждый момент времени мы располагаем полным списком PST, построенных из гипотез с наивысшей оценкой. Алгоритм останавливается, когда получено заданное количество PST с лучшей оценкой или по исчерпании списка структур. Таким образом, мы получаем заданное число PST заданной длины, и избегаем нахождения и оценки всех возможных вариантов PST для данного спектра.

Глава 4. Программный комплекс Proteos

Алгоритм CrystalTag реализован в составе программного комплекса, получившего название Proteos. Этот программный комплекс обеспечивает полный цикл интерпретации масс-спектрометрических данных с использованием методологии поиска PST, начиная с чтения файлов исходных данных и заканчивая формированием биохимически значимой гипотезы о составе исходной смеси белков.

Данный комплекс реализует стратегию устойчивого к ошибкам поиска PST в белковых базах данных, предложенную в [2]. Результатом работы комплекса является ранжированный список белков-кандидатов, присутствие которых в составе исходной смеси наиболее вероятно. Ранжирование проводится на основании методики, описанной в [4], подразумевающей вероятностную оценку совпадения набора PST и аминокислотных последовательностей белков. Вероятностная оценка для каждого белка строится как вероятность случайного совпадения компонент PST и аминокислотной последовательности белка, представленного как последовательность символов с пуассоновой характеристикой

появления. Оценка строится с учетом среднего содержания аминокислот в белковых последовательностях и распределения по массам триптических и нетриптических пептидов.

Выходной список белков-кандидатов сгруппирован по принципу возможной гомологии белков. У гомологичных белков значительная часть аминокислотной последовательности может совпадать и, поэтому, если в списке белков-кандидатов появляется ряд гомологичных белков, из этого ряда имеет смысл рассматривать только белок с наивысшим рейтингом. Группировка белков позволяет быстро выполнить эту задачу.

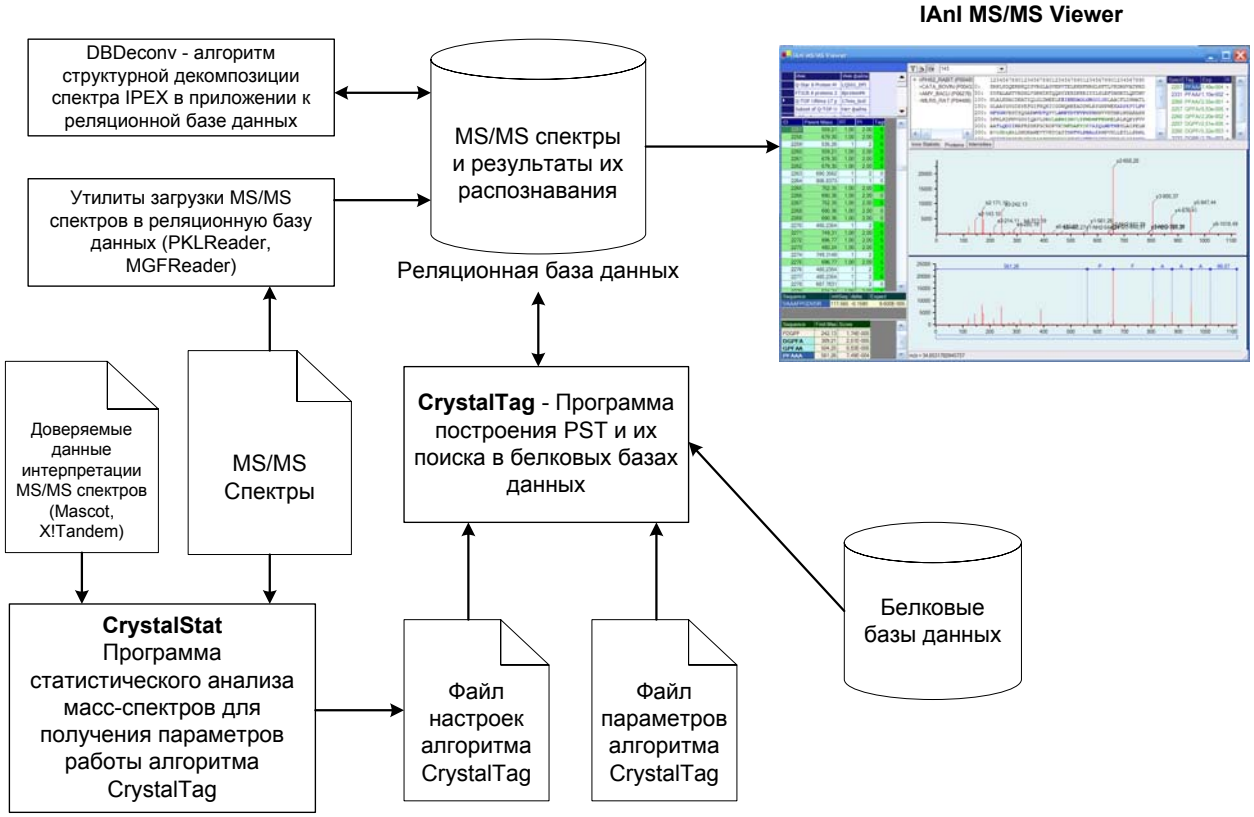

Рис. 4. *Общая структура программного комплекса Proteos*

На момент написания данной работы программный комплекс Proteos реализован на уровне инженерного прототипа и доступен в лаборатории биомедицинской масс-спектрометрии Института аналитического приборостроения РАН.

Общая структура программного комплекса Proteos представлена на рисунке 4. Программный комплекс включает в себя программные средства для проведения полного цикла интерпретации данных ВЭЖХ-МС/МС экспериментов, включающего следующие этапы:

1 этап. Подготовку данных для работы алгоритмов биохимической интерпретации. Этот этап состоит из следующих стадий:

Ввод набора МС/МС спектров в базу данных программного комплекса – осуществляется программами PKLReader и MGFReader

Извлечение из набора МС/МС спектров аналитически значимой информации – раскрытие изотопных и зарядных распределений, представленных в масс-спектрах – выполняется утилитой DBDeconv, предоставляющую реализацию алгоритма IPEX [5], адаптированную для работы с реляционной базой данных.

Статистический анализ спектров с привлечением доверяемых результатов интерпретации масс-спектров с целью получения набора оценок критериев, используемых для интерпретации масс-спектров – выполняется программой CrystalStat.

2 этап. Собственно интерпретация данных масс-спектрометрического протеомного эксперимента и формирование итогового списка белков-кандидатов. Этот этап полностью

выполняется программой CrystalTag. Результат интерпретации сохраняется в реляционной базе данных комплекса. Для своей работы программа CrystalTag использует масс-спектры, предварительно сохраненные в реляционной базе данных, белковые базы данных, представленные в текстовом формате FASTA, текстовый файл настроек алгоритма, сгенерированный программой CrystalStat, и текстовый файл параметров работы алгоритма, предоставленный пользователем программы.

3 этап. Визуализация данных масс-спектрометрического эксперимента и результатов его интерпретации. Выполняется приложением IAnI MS/MS Viewer на основании данных, сохраненных в базе данных.

Программный комплекс Proteos в данной работе выполняет функцию стенда, который служит для испытания и характеризации алгоритма CrystalTag.

Глава 5 Характеризация алгоритма CryatalTag

Тестирование алгоритма

Тестирование алгоритма CrystalTag выполнялось на ранее описанных выборках данных. Основной целью тестирования было оценить производительность алгоритма и качество распознавания PST. Для оценки производительности использовался полный набор масс-спектров. Для оценки качества распознавания PST из полного набора масс-спектров были отобраны только спектры известных пептидов, содержащие последовательности сигналов ионов основных серий, достаточно длинные для построения PST. Тестирование выполнялось на рабочей станции, оснащенной процессором Intel Pentium M 1.7 GHz и 1 GB оперативной памяти.

Среднее время работы алгоритма для всех протестированных вариантов осталось в субмилисекундном диапазоне. Время выполнения алгоритма для отдельного спектра варьируется от 0.08 мсек. до 24.3 мсек. Установлено, что за 3 мсек. алгоритм в 96% случаев успевает найти хотя бы один верный PST, если таковой существует. Наибольшее время занимает обработка спектров, содержащих множество слабых, неинформативных сигналов. Это заставляет рекомендовать ограничение по числу сигналов масс-спектра в 100-150 наиболее сильных сигналов. Благодаря высокому быстродействию, алгоритм CrystalTag можно рекомендовать для использования в системах DDA (Data Dependent Acquisition), требующих быстрой оценки качества спектра во время работы масс-спектрометра.

***Таблица 2**. Результаты тестирования алгоритма CrystalTag*

	Q-TOF Ultima (Выборка №1)			Bruker Esquire (Выборка №2)			Finnigan LTQ-FT (Выборка №3)		
Число PST/ Длина PST	5/5	20/5	5/4	5/5	20/5	5/4	5/5	20/5	5/4
Всего спектров	1382	1382	1382	6048	6048	6048	10000	10000	10000
Спектров, пригодных для тестирования	213	213	227	983	983	1186	7356	7356	8845
Время работы алгоритма	0.55	0.64	0.43	0.51	0.85	0.26	0.47	0.71	0.41
% корректно распознанных PST	90.2%	95.8%	94.8%	70.8%	82.8%	73.9%	97.8%	98.4%	98.1%

Результаты тестирования качества распознавания PST закономерно соотносятся с аналитическими характеристиками приборов, на которых были получены масс-спектры. Чем выше аналитические характеристики приборов, тем больший процент PST удается корректно опознать.

Эффект использования алгоритма CrystalTag

Эффект использования метода устойчивого к ошибкам поиска в базах данных на основании PST продемонстрирован на примере выборки масс-спектров №1, для которой известен точный состав белков в пробе для анализа. Диаграмма на рис. 5 показывает, что кроме масс-спектров триптических пептидов, обнаруживаемых при помощи систем идентификации белков методом картирования фрагментов, в массиве спектров обнаруживается значительное количество масс-спектров пептидов (до 40% общего количества), несущих пост-трансляционные модификации, а также полученных в результате неполного или неспецифического гидролиза.

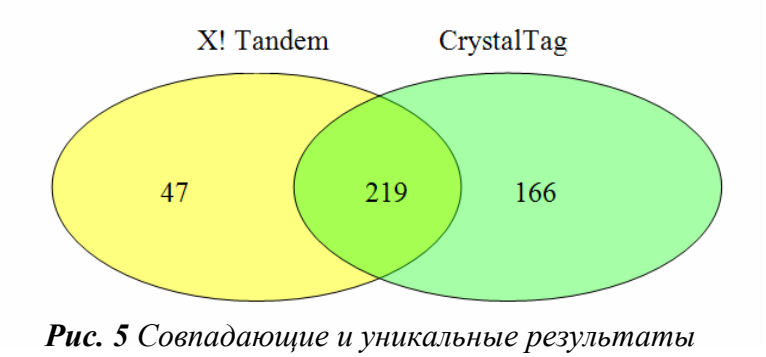

Рис. 5 *Совпадающие и уникальные результаты идентификации пептидов для X!Tandem и CrystalTag*

Обычным эффектом вовлечения в результаты идентификации таких пептидов является увеличение покрытия аминокислотной последовательности белка идентифицированными пептидами. В результатах обработки выборок данных, описанных в работе, наблюдалось увеличение покрытия последовательности белка до 1.5 раз по сравнению с результатами обработки выборок масс-спектров при помощи систем идентификации белков Mascot и X!Tandem.

Сопоставление с существующими аналогами

В качестве наиболее адекватного образца для сопоставления были выбраны программы Inspect и Pepnovo разработанные в Университете Южной Калифорнии под руководством проф. П. А. Певзнера [6]. Эти разработки проводились параллельно с данной работой и имеют сходную с данной работой идеологию обработки массивов масс-спектров. Inspect предназначен для выявления пост-трансляционных модификаций пептидов за счет реализации стратегии устойчивого к ошибкам поиска в базах данных. Pepnovo – это программа совмещающая распознавание de novo и поиск Peptide Sequence Tag во фрагментных масс-спектрах.

Анализ быстродействия производился на основе внедрения в исходный код Inspect и Pepnovo закладок для измерения чистого времени работы алгоритма. Состав выборок масс-спектров и параметров алгоритмов для тестирования определялся возможностями настройки Pepnovo и Inspect на момент исследования (декабрь 2006 г.).

Таблица 3. *Сопоставление алгоритмов Inspect, PepNovo и CrystalTag*

	QTOF Ultima (Выборка №1)	Bruker Esquire (Выборка №2)			Finnigan LTQ-FT (Выборка №3)		
Число PST /Длинна PST	5/4	5/5	20/5	5/4	5/5	20/5	5/4
Качество распознавания (CrystalTag)	94.8%	70.8%	82.8%	73.9%	97.8%	98.4%	98.1%
Качество распознавания (PepNovo)		66.1 %	80.0%	69.7%	88.2%	96.4%	88.7%
Качество распознавания (Inspect)	79.8%			65.8%			
Скорость работы (CrystalTag) мсек.	0.43	0.51	0.85	0.26	0.47	0.71	0.41
Скорость работы (PepNovo) мсек.		82	89	78	12	14	11
Скорость работы (Inspect) мсек.	0.94			0.72			

Сопоставление показало, что алгоритм, включенный в состав Inspect, уступает как по скорости, так и по качеству распознавания. PepNovo показал качество распознавания, сравнимое с результатами работы CrystalTag, однако сильно проиграл по производительности.

Итак, алгоритм CrystallTag показал себя самым быстрым методом распознавания PST. Что касается качества распознавания, то следует отметить, что алгоритм PepNovo имеет гораздо более изысканную процедуру оценки PST нежели CrystallTag. Тем не менее, качество распознавания остается на том же уровне. Возможно, это свидетельствует о достижении некоторого предела метода, связанного не столько с качеством алгоритма, сколько с природой метода.

Применение программного комплекса Proteos к данным актуальных биологических исследований

В сотрудничестве с коллективом лаборатории протеомики ИБХ РАН удалось применить алгоритмы и методы, реализованные в составе комплекса Proteos, к данным актуальных протеомных исследований.

Повышение достоверности идентификации белков*:* Для надежной идентификации белка необходимо указать не менее 2 пептидов, уникальных для данного белка. При исследовании протеома митохондрий сердца быка были идентифицированы более 500 белков. Однако из них около 250 белков были идентифицированы не более, чем по единственному уникальному пептиду. Объектом внимания стали ~100 белков, для которых обнаружен единственный пептид и не существует пептидов совпадающих с пептидами других белков. Из общей выборки в 117 422 спектра были выделены 25 спектров, которые потенциально соответствуют дополнительным пептидам 19 белков.

Восстановление последовательности трансмембранных участков белков: В ИБХ РАН предложена методика пробоподготовки, позволяющая получить пробы пептидов трансмембранных участков белков. Масс-спектры таких нетриптических пептидов с трудом поддаются интерпретации стандартными средствами. При исследовании результата LC-MS/MS эксперимента над препаратом мембраны митохондрий клеток бычьего сердца удалось обнаружить спектры 18 пептидов, которые удается идентифицированных как потенциальные трансмембранные домены 11 белков.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Разработан новый высокопроизводительный метод восстановления пептидной аминокислотной последовательности, представленной во фрагментном масс-спектре.
2. Разработана методика численной оценки эмпирических критериев на основе использования статистической информации о фрагментных масс-спектрах пептидов.
3. Разработан высокопроизводительный адаптивный алгоритм распознавания аминокислотной последовательности пептида во фрагментном масс-спектре, названный CrystalTag.
4. На основе предложенной методики численной оценки критериев разработана процедура автоматической настройки параметров алгоритма CrystalTag.
5. Предложенный метод реализован как набор программных компонент в составе программного комплекса, выполняющего полный цикл интерпретации данных масс-спектрометрического эксперимента в белковых анализах.
6. Показана универсальность метода для данных, полученных на масс-спектрометрах различной архитектуры.
7. Проведен сравнительный анализ параметров разработанного программного комплекса с существующими аналогами, который показал преимущества по производительности и качеству распознавания аминокислотных последовательностей пептидов

ЦИТИРУЕМАЯ ЛИТЕРАТУРА:

1. Roepstorff P., Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides // *Biomed. Mass Spectrom.* - 1984 Nov, vol. 11(11), p. 601.
2. Mann M., Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags //*Anal. Chem.* 1994. vol. 66(24) pp. 4390-4399.
3. Bartels C. Fast algorithm for peptide sequencing by mass spectrometry // *Biomed. Environ. Mass Spectrom.* 1990. vol. 19, pp. 363-368
4. Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. // *Anal. Chem*. 2003. vol. 75(6), pp. 1307-1315.
5. Макаров В.В., Савельев С.К., Лютвинский Я.И., Веренчиков А.Н., Краснов Н.В. Алгоритм извлечения аналитически значимой информации из масс-спектрометрических данных экспериментов протеомики. // *Научное приборостроение*. – 2006. - т.16. №2, сс. 92-100.
6. Frank A, Tanner S, Bafna V, Pevzner P. Peptide sequence tags for fast database search in mass-spectrometry. // *J. Proteome Res.* 2005. vol. 4(4), pp. 1287-1295.

ПУБЛИКАЦИИ, ОТРАЖАЮЩИЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ:

1. Лютвинский Я.И., Краснов Н.В. Разработка CRYSTALTAG – алгоритма частичного распознавания фрагментных масс-спектров пептидов. Научное приборостроение. 2005. Т.15, №3 С.108-113
2. Лютвинский Я.И., Макаров В.В., Краснов Н.В., Подольская Е.П., Веренчиков А.Н. Частичная расшифровка аминокислотной последовательности пептида по его фрагментному масс-спектру: алгоритм и результаты применения. Научное приборостроение. 2006. Т.16, №3 С.122-131
3. Лютвинский Я.И., Макаров В.В., Веренчиков А.Н. Использование статистики фрагментации ионов для частичной интерпретации фрагментных масс-спектров пептидов. Всероссийская конференция «Масс-спектрометрия и ее прикладные проблемы», г. Москва, 2005 г.
4. Лютвинский Я.И., Макаров В.В., Краснов Н.В. Crystaltag – новый алгоритм частичной интерпретации масс-спектров пептидов. Тезисы докладов III съезда общества биотехнологов им. Ю.А. Овчинникова. Москва, 25-27 октября 2005 г.
5. Лютвинский Я.И., Новиков А.В., Федорова Г.А. Фрагментация пептидов в источнике электроспрей как способ извлечения информации о первичной структуре пептида на масс-спектрометре MX5305. Тезисы докладов III съезда общества биотехнологов им. Ю.А. Овчинникова. Москва, 25-27 октября 2005 г.
6. Лютвинский Я.И., Макаров В.В., Краснов Н.В. Использование статистики фрагментации ионов для частичного распознования масс-спектров пептидов. Тезисы докладов конференции «Аналитическое приборостроение» С.Петербург. 2005.