*Биомедицинская химия, 2014 том 60, вып. 6, с. 707-712.*

УДК 577.332


# ПРОСТОЙ МЕТОД ОЦЕНКИ ВЕРОЯТНОСТИ ДЕТЕКЦИИ ПЕПТИДА ПРИ МАСС-СПЕКТРОМЕТРИИ С ЭЛЕКТРОСПРЕЙНОЙ ИОНИЗАЦИЕЙ

***А.В. Рыбина, В.С. Скворцов*, А.Т. Копылов, В.Г. Згода***

Институт биомедицинской химии им. В.Н. Ореховича,
ул. Погодинская, 10, 119121, Москва; эл. почта: vladlen@ibmh.msk.su

Рассмотрена возможность отбора перспективных пептидов, использующихся для детекции и количественного определения белков в направленной масс-спектрометрии методом положительной электроспрейной ионизации. В основе данного метода лежит предсказание интенсивности масс-спектрометрических пиков с использованием линейно-регрессионной модели. Данный метод имеет априорные ограничения: предварительный отбор пептидов должен быть организован таким образом, чтобы при pH=2,5 выбранные пептиды должны быть в основном представлены ионами 2+ и 3+. Рассматриваются только пептиды, имеющие на С-конце остатки аргинина или лизина. В качестве независимых переменных используются аминокислотный состав пептида, концентрация пептида и отношение площадей полярных поверхностей полярных объёмов к общим. Из нескольких рассмотренных комбинаций переменных лучшая линейно-регрессионная модель имеет коэффициент детерминации при скользящем контроле 0,54. Модель уверено отличает пептиды с высокой степенью ответа от пептидов с низкой степенью ответа, и если отобранных по простым критериям пептидов слишком много, позволяет отобрать только самые перспективные. Данный способ фильтрации, простой и быстрый, может быть с успехом применён для сокращения списка наблюдаемых пептидов.

**Ключевые слова:** пептид, масс-спектрометрия, электроспрейная ионизация, предсказание свойств.

## ВВЕДЕНИЕ

В направленной масс-спектрометрии существует проблема: исследователь должен знать точно те пептиды, по которым он собирается отслеживать наличие (а в идеале и количество) интересующего его белка. В то же время, различия физико-химических свойств отдельных пептидов напрямую влияют на возможность их детекции, а также на величину интенсивности пиков на масс-спектрограммах. Проблема отбора таких пептидов чрезвычайно актуальна, ведь формирование потока ионов из заряженной капли в электроспрее – сложный процесс, обусловленный многими факторами. Данная проблема достаточно подробно исследовалась в конце 80-х начале 90-х годов прошлого века [1, 2] и в ряде работ анализировалась возможность отбора хорошо детектируемых пептидов [3-7]. В частности, Fusaro с соавторами [7] предсказывали пептиды с высоким уровнем ответа на основании анализа более 30 дескрипторов, описывающих свойства пептидов методом “случайного леса”. Для подобных предсказаний чаще всего и используются классификационные методы. К сожалению, селективность этих методов невысока, а наилучшие результаты даёт самый простой: наличие в базах данных “PeptideAtlas”, “MRMatlas” или любых других указания, что данный пептид наблюдали в эксперименте ранее.

Главное ограничение на выбор пептида накладывает выбранный метод ионизации. Пептиды, как правило, ионизируют путём протонирования. Если мы наблюдаем за ионами, имеющими заряд 2+, то желательно, чтобы при установленном в эксперименте pH большая часть пептида присутствовала в виде данного иона. Если возможно образование ионов 3+ или 4+, то необходимо учитывать и их. Задача расчёта ионных форм – задача более

* - адресат для переписки

сложная, чем принято считать. В какой-то степени её можно обойти, проводя эксперимент при очень кислых значениях pH, когда наиболее предпочтителен ион в предельно протонированной форме. В данной работе мы использовали pH=2,5. К сожалению, зависимость ионизации пептидов от температуры в литературе широко не рассматривается. Тем не менее, первичный отбор пептидов не сложно провести, если известны параметры ионизации. Возникает вопрос: возможно ли, проведя первичную селекцию пептидов отобрать лучшие из них методами прямого предсказания величины интенсивности пиков при масс-спектрометрии? Результаты данной работы показывают, что это возможно и весьма простым методом.

## МЕТОДИКА

Синтез пептидов, имитирующих триптические, был описан ранее [8]. Всего в работе было использовано 568 синтезированных пептидов, разделённых на 8 проб (см. приложение 1). Хроматографическое разделение пептидов проводили на колонке с обращённой фазой Zorbax (C18 eclips 3×100 мм) (“Agilent”, США) на хроматографе Agilent 1260 (“Agilent”). Скорость потока растворителя 0,2 мл/мин. Градиент: 5% – 50% B (A – вода с 0,1% муравьиной кислотой, B – водный 80% ацетонитрил с 0,1% муравьиной кислотой) 40 мин и 45% – 100% B 5 мин. Детектирование пептидов проводили на времяпролетном масс-спектрометре Agilent 6550. Масс-спектрометрический анализ проводили при при температуре капилляра 250°C и разнице потенциалов ионизации 3500V. Для каждой из проб проводили по 3 технических повтора измерений. Полученная в результате выборка данных подробно описана в приложении 1.

Идентификацию пептидов, определение площади под пиком и интенсивность проводили в автоматизированном режиме с использованием программного обеспечения Progenesis (“Nonlinear Dynamics”, США) [9]. В качестве меры интенсивности пика использовали величину “Normalized abundance” (В дальнейшем её будем называть просто интенсивность). При окончательной фильтрации применялись следующие правила:

1. При идентификации пептидов по величине *m/z* (масса/заряд) использовали только данные для ионов 2+ и 3+.

2. Максимальная дельта масс при поиске была установлена для ионов 2+ – не более 0,025, для 3+ – не более 0,018.

3. При наличии обоих ионов нормализованная величина “abundance” суммировалась. В каждом случае и для каждого иона проводилось усреднение по результатам 3-х экспериментов.

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Среди идентифицированных 387 пептидов, оба варианта ионов были обнаружены только в 102 случаях. Существенная часть неидентифицированных пептидов, по-видимому, может быть выявлена при применении других методов обработки данных. Однако, для воспроизводимости результатов было решено ограничиться автоматизированной процедурой.

Следует отметить, что величины интенсивности (I), молекулярного веса (MW) и концентрации (C, пмоль/мкл) сильно коррелируют между собой (коэффициенты корреляции больше 0,6). Если связь между I и C собственно и даёт возможность количественно определять пептид масс-спектрометрическим способом (рис. 1), то корреляция между C и MW носит случайный характер и связана с особенностями приготовления проб. Используя зависимость между I и C, можно рассчитать значение ожидаемой величины интенсивности и использовать в дальнейшем в качестве зависимой переменной следующую величину:

I = log(I) - log(Iожидаемая)

Распределение полученной таким образом величины близко к нормальному (рис. 2). Если величина положительная, то, согласно нулевой гипотезе пептид детектируется лучше ожидаемого, если отрицательная – хуже. Собственно нулевая гипотеза – интенсивность масс-спектрометрического пика для пептида, отобранного при заданных ограничениях (усреднённый заряд пептида при pH = 2,5 не меньше +2) прямо пропорциональна концентрации (в логарифмической шкале).

Самый простой вариант для определения независимых переменных в данном случае это аминокислотный состав. Соответственно, для каждого пептида был рассчитан вектор из 20 значений, каждое из которых было числом встречаемости одного из аминокислотных остатков в пептиде. Уравнения зависимости были построены с использованием линейной регрессии. Начальная модель имела слабые характеристики ($R^2$=0,35; средняя ошибка MAE=0,18; среднеквадратичная ошибка MSE=0,07; максимальная ошибка ME=1,32). Однако, тест по случайному смешиванию зависимых переменных показал, что модель значимая (величины усреднены по 10 попыткам: $R^2$=0,05; MAE=0,25; MSE=0,1; ME=1,48).

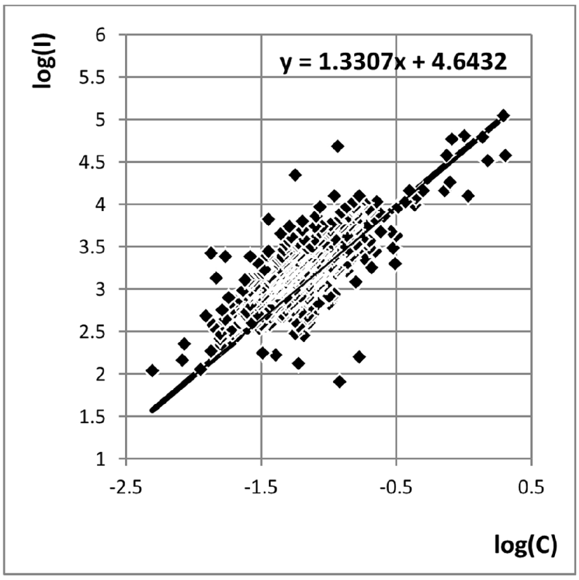


**Рисунок 1.** Зависимость суммарной интенсивности пика от концентрации пептида.

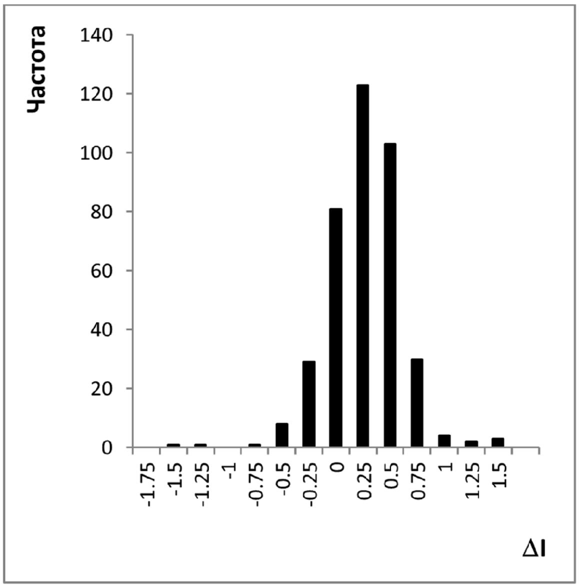


**Рисунок 2.** Распределение величины отклонения интенсивности пиков от ожидаемой для 387 идентифицированных пептидов.

*A priory* можно предположить наличие в выборках некоторого числа резко выделяющихся наблюдений. Они могут быть связанны как с уникальностью пептида, так и возможными ошибками при идентификации пептидов. Для выявления таких наблюдений можно использовать робастные подходы. В нашем случае мы использовали программу для построения нейронных сетей обратного распространения [10], которая, как частный случай, может рассматривать и линейные модели. Результаты робастного моделирования (рис. 3) явным образом указывают на наличие 6 случаев достоверных отклонений, которые можно исключить (наиболее вероятно, что это ошибки идентификации пептидов). В результате параметры линейно-регрессионной модели на основе только аминокислотного состава существенно улучшились ($R^2$=0,48; MAE=0,17; MSE=0,05; ME=0,94; в случайном тесте: $R^2$=0,06; MAE=0,23; MSE=0,09; ME=1,47).

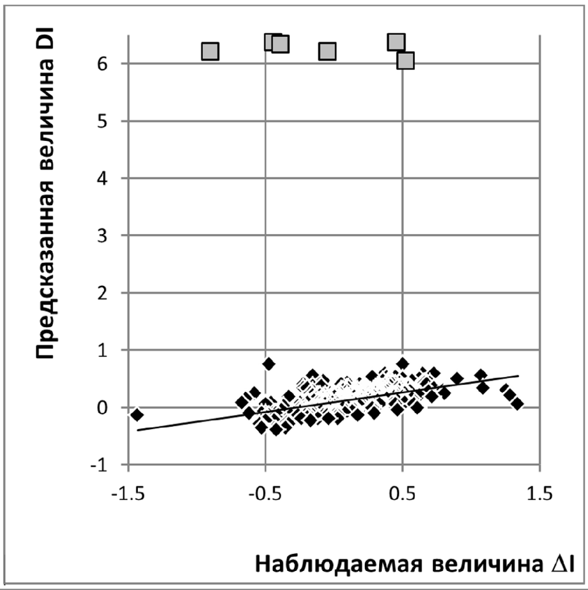


**Рисунок 3.** Тест на наличие выбросов с использованием робастной модели.

Для повышения предсказательной силы модели были построены несколько дополнительных уравнений, когда к аминокислотному составу добавляли различные расчётные параметры пептидов (заряд в заданных условиях, оценка гидрофобности, дескрипторы формы и др.). За небольшим исключением это не дало значимого улучшения предсказательной силы. Существенный для повышения предсказательной силы модели вклад вносит логарифм концентрации (log(C)). Сложно сказать, связано ли это с неточностью функции расчёта ожидаемой интенсивности или несёт физический смысл. Из общих соображений можно предположить, что количество ионов, вылетающих из капли спрея, может иметь более сложную зависимость от количества вещества в капле, чем линейную. Впрочем, учитывая, что наблюдения не имеют характер равномерного распределения, способ расчёта калибровочной кривой для зависимости интенсивности от концентрации также может вносить свои поправки. Два дополнительных параметра, использованных в конечном варианте модели: отношение полярной поверхности к общей поверхности и отношение полярного объёма к общему объёму, к сожалению, требуют моделирования трёхмерной структуры пептида и в некоторой степени зависят от выбора конформации. Проблема требует дальнейшего исследования, но учитывая, что $R^2$ в конечной модели с этими параметрами всего на 0,05 лучше, то можно считать, что, несмотря на то, что улучшение значимое, в данном случае им можно и пренебречь в пользу простоты расчётов. Стоит отметить также, что вклад ряда аминокислотных остатков несущественен и из уравнения конечной модели они исключены (таблица). В случае лизина и аргинина это объясняется достаточно просто. Так как синтезированные пептиды (прил. 1) имитировали пептиды, полученные в результате трипсинолиза, то все они заканчивались на “R” (аргинина) или “K” (лизина). Данный факт определяет одно из основных ограничений для области применимости модели: предсказания адекватны только для пептидов, имеющие на С-конце эти аминокислотные остатки.

Лучшая модель (таблица, рис. 4), полученная в рамках нашей работы имела следующие параметры: $R^2$=0,58; MAE=0,15; MSE=0,04; ME=0,72; в случайном тесте: $R^2$=0,04; MAE=0,24; MSE=0,09; ME=1,53. Модель показала хорошую устойчивость в тестах по кроссвалидации. Использовались два варианта – скользящий контроль методом выбрасыванием по одному ($R^2$=0,54) (рис. 4В) и делением выборки пополам ($R^2$= 0,55 и 0,53) (рис. 5). В последнем случае выборка сортировалась по значению I и делилась на чётные и нечётные номера.

Несомненно, коэффициент детерминации меньше 0,6 не так уж и хорош, тем не менее, данная модель уверено отличает пептиды с высокой степенью ответа от пептидов с низкой степенью. Причём, исследователь может устанавливать величину концентрации либо равной для всех пептидов-кандидатов, либо разной по своим собственным соображениям. Напомним, что данный тест является вторичным, пептиды уже отбирались таким образом, чтобы масс-спектрометрия с электроспрейной ионизацией методом протонирования могла обнаружить ионы 2+ и 3+ при заданном в эксперименте pH. В случае, если отобранных по простым критериям пептидов слишком много, данный способ фильтрации, простой и быстрый, может быть с успехом применён для сокращения списка наблюдаемых пептидов. Необходимо также обратить внимания на результаты проверки выборки на выбросы с использованием робастных моделей. Конечно, эта проблема

*Таблица.* Коэффициенты линейного уравнения с наилучшими параметрами, включающего аминокислотный спектр пептида, концентрацию и параметры формы.

| log(C) | A | V | I | L | P | S | C | D | Q | H | F | W | PSA/S | PV/V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -0,341 | 0,055 | 0,203 | 0,163 | 0,355 | 0,069 | -0,054 | -0,061 | -0,171 | -0,109 | -0,089 | 0,188 | -0,049 | -0,185 | -0,117 |

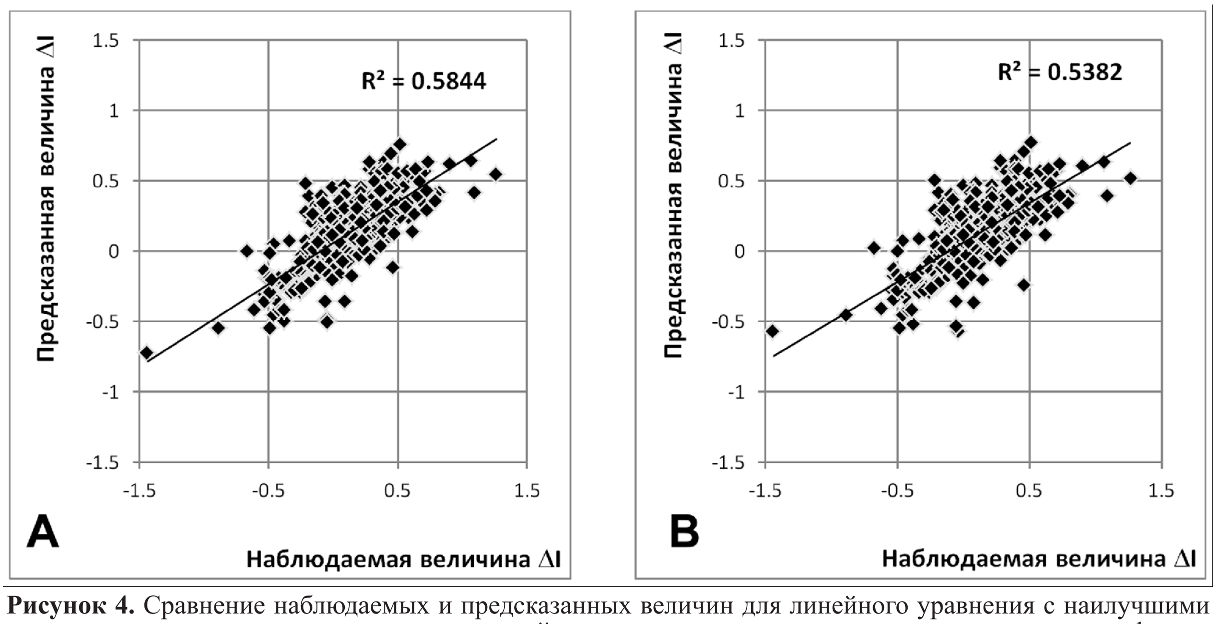


**Рисунок 4.** Сравнение наблюдаемых и предсказанных величин для линейного уравнения с наилучшими параметрами, включающего аминокислотный спектр пептида, концентрацию и параметры формы (таблица). **А**. Обучение. **В**. Проверка методом выбрасывания по одному (кроссвалидация).

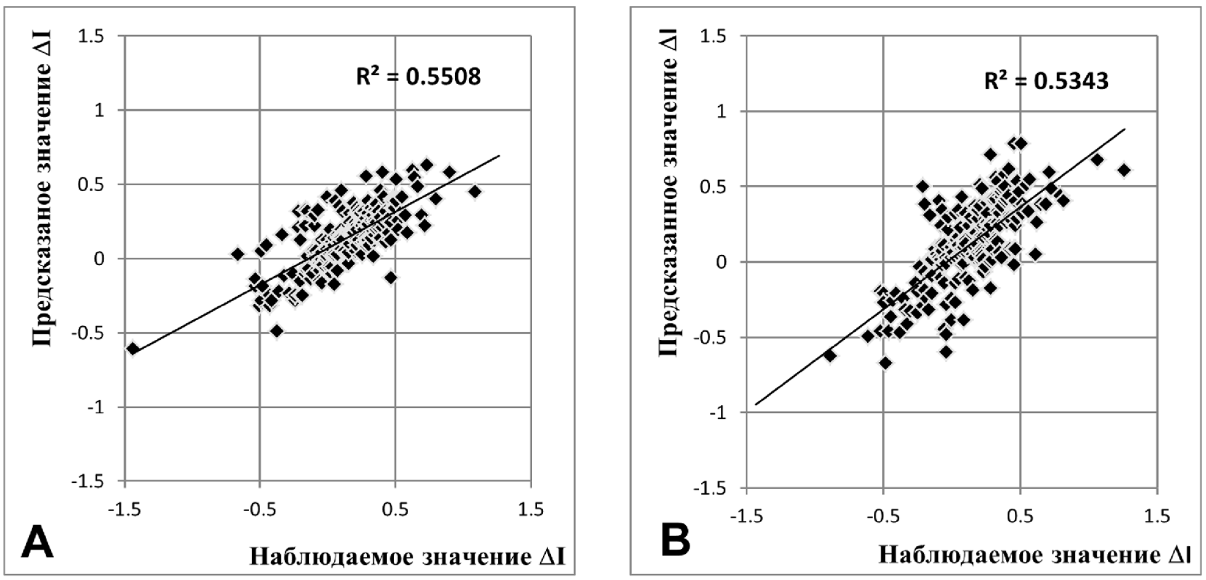


**Рисунок 5.** Сравнение наблюдаемых и предсказанных величин в тесте деления выборки пополам. **A**. Нечётные номера. **B**. Чётные номера.

требует дальнейшего анализа, но ошибки при идентификации пептидов по масс-спектрам не так уж редки. Использование же робастных моделей может позволить выявить те пептиды, вероятность неправильной идентификации которых, наиболее велика.

*Работа выполнена в рамках государственного задания по программе фундаментальных исследований Государственных академий. Масс-спектрометрические измерения выполнялись в рамках работ, поддержанных грантом РНФ № 14-25-00132.*

Дополнительные материалы и приложения свободно доступны в электронной версии статьи на сайте журнала.

## ЛИТЕРАТУРА


1. *Guevremont R., Siu K.W.M., Le Blanc J.C.Y., Berman S.S.* (1992) J. Am. Soc. Mass Spectrom., **3**, 216-224. DOI:10.1016/1044-0305(92)87005-J
2. *Fenn J.B.* (1993) J. Am. Soc. Mass Spectrom., **4**, 524-535. DOI:10.1016/1044-0305(93)85014-O.
3. *Braisted J.C., Kuntumalla S., Vogel C., Marcotte E.M., Rodrigues A.R., Wang R., Pieper R.* (2008) BMC Bioinformatics, **9**, 529. DOI:10.1186/1471-2105-9-529.
4. *Jin S.* (2007) Integrated Data Modeling in High-throughput Proteomices, Doctoral dissertation, Washington State University. ISBN: 978-0-549-47388-6.
5. *Noyce A.B., Smith R., Dalgleish J., Taylor R.M., Erb K.C., Okuda N., Prince, J.T.* (2013) J. Proteome Res., **12**, 5742-5749. DOI:10.1021/pr400727e.
6. *Schliekelman P., Liu S.* (2013) J. Proteome Res., **13**, 348-361. DOI:10.1021/pr400034z.
7. *Fusaro V.A., Mani D.R., Mesirov J.P., Carr S.A.* (2009) Nat. Biotechnol., **27**, 190-198. DOI: 10.1038/nbt.1524.
8. *Moshkovskii S.A., Zgoda V.G, Sokolov A.S., Mazur A.M., Prokhortchouck E.B., Skryabin K.G., Ilina E.N., Kostryukova E.S., Alexeev D.G., Tyakht A.V., Gorbachev A.Y., Govorun V.M., Archakov A.I.* (2014) J. Proteome Res., **13**, 183-190. DOI:10.1021/pr400883x.
9. Progenesis v.2.6 software (NonLinear Dynamics Ltd, Newcastle, USA).
10. *Belkina N.V., Krepets V.V., Shakin, V.V.* (2002) Autom. Remote Control, **63**, 66-75. DOI:10.1023/A:1013783319376.


Поступила: 05. 10. 2014.

## A PLAIN METHOD OF PREDICTION OF VISIBILITY OF PEPTIDES IN MASS SPECTROMETRY WITH ELECTROSPRAY IONIZATION

***A.V. Rybina, V.S. Skvortsov*, A.T. Kopylov, V.G. Zgoda***

Institute of Biomedical Chemistry,
10, Pogodinskaya str., Moscow, 119121 Russia; e-mail: vladlen@ibmh.msk.su

A new method for screening of essential peptides for protein detection and quantification analysis in the direct positive electrospray mass spectrometry has been proposed. Our method is based on the prediction of the normalized abundance of the mass spectrometric peaks using a linear regression model. This method has the folowing limitations: (i) selected peptides should be taken so that at pH 2.5 the tested peptides must be presented mainly as the 2+ and 3+ ions; (ii) only peptides having C-terminal lysine or arginine residues are considered. The amino acid composition of the peptide, the peptide concentration, the ratio of the polar surface of peptide to common surface and ratio of the polar volume to common volume are used as independent variables in equation. Several combinations of variables were considered and the best linear regression model had a determination coefficient in leave-one-out validation procedure equal 0.54. This model confidently discriminates peptides with high response ability and peptides with low response ability, and therefore it allows to select only the most promising peptides. This screening method, a plain and fast, can be successfully applied to reduce the list of observed peptides.

**Key words:** peptide, mass-spectrometry, electrospray ionization, property prediction.