УДК 573.6
DOI 10.25205/1818-7900-2020-18-2-31-42

BioNet: моделирование масс-спектров пептидов

Р. Ю. Епифанов [1], **Д. А. Афонников** [1, 2]

[1] *Новосибирский государственный университет*
Новосибирск, Россия

[2] *Институт цитологии и генетики СО РАН*
Новосибирск, Россия

Аннотация

Определение белкового состава живой клетки (протеома) – одна из важнейших задач современной биологии. Универсальным инструментом для исследования протеома является масс-спектроскопия. Расшифровка масс-спектров является сложной задачей, так как не до конца известны механизмы диссоциации белков в экспериментальных установках, а также влияние совокупности внешних факторов на данный процесс. Для совершенствования существующих или разработки новых алгоритмов расшифровки масс-спектров требуется большое количество данных по аннотированным масс-спектрам пептидов с известной последовательностью. В статье описана разработка алгоритма *in silico* моделирования масс-спектра пептидов, решающего проблему учета влияния неканонического аминокислотного состава и посттрансляционных модификаций на процесс диссоциации. Для проверки работоспособности построенного алгоритма проведено сравнение его эффективности с аналогами. Показано, что точность предложенного метода выше, особенно для пептидов, подверженных посттрансляционным модификациям.

Ключевые слова

in silico масс-спектроскопия, нейросетевые методы, неканонические аминокислоты, посттрансляционные модификации

Благодарности

Работа поддержана грантами РФФИ № 17-00-00470 (К), 17-00-00462

Для цитирования

Епифанов Р. Ю., Афонников Д. А. BioNet: моделирование масс-спектров пептидов // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, № 2. С. 31–42. DOI 10.25205/1818-7900-2020-18-2-31-42

BioNet: Peptide Mass-Spectrum Prediction

R. Yu. Epifanov [1], **D. A. Afonnikov** [1, 2]

[1] *Novosibirsk State University*
Novosibirsk, Russian Federation

[2] *Institute of Cytology and Genetics SB RAS*
Novosibirsk, Russian Federation

Abstact

The importance of the biological properties of proteins to cells cause actively exploring their amino acid composition (primary structure). The versatile tool of cell proteome exploring is mass-spectroscopy. The interpretation of mass-spectroscopy data is complex challenge because it remains uncertain peptide dissociation mechanisms and external factor influence to peptide fragmentation process. Moreover, a lot of mass-spectroscopy data is required to enhancement existing or development novel algorithms to interpret peptide mass-spectra. The article describes development of algorithm for *in silico* generation peptide mass-spectra covered the problem of influence noncanonical amino acid composition and posttranslational modifications to dissociation process. Developed algorithm was compared with analogues and evaluated over experimental data.

Keywords

in silico mass-spectroscopy, artificial neural networks, noncanonical amino acids, posttranslational modifications.

Acknowledgements

The research was supported by RFBR grants no. 17-00-00470(K), 17-00-00462

For citation

Epifanov R. Yu., Afonnikov D. A. BioNet: Peptide Mass-Spectrum Prediction. *Vestnik NSU. Series: Information Technologies*, 2020, vol. 18, no. 2, p. 31–42. (in Russ.) DOI 10.25205/1818-7900-2020-18-2-31-42

Введение

Белки являются важнейшими биологическими объектами. Разнообразие их функции в клетках организмов выше, чем у других биополимеров. В зависимости от выполняемых функций белки подразделяют на группы: ферменты, которые участвуют в специфическом катализе биологических реакций; структурные белки, которые определяют структуру тканей и форму тела животных; транспортные белки, участвующие в переносе веществ из клетки в клетку, и др.

Белки являются полимерами и состоят из аминокислот, связанных пептидными связями. Первичная структура белка может быть представлена в виде строки символов, каждый из которых представляет одну из 20 канонических аминокислот. Левый конец последовательности называется N-терминальным, правый – C-терминальным. Определение последовательности белка (его первичной структуры) является ключевым этапом в понимании механизмов его функционирования. Исследование первичной структуры белков возможно с помощью методов геномики. Секвенирование генома организма позволяет производить определение первичной структуры белка по последовательности кодирующих его генов [1]. Недостатком такого метода является то, что с помощью него невозможно определить наличие посттрансляционных модификаций в первичной структуре, которые оказывают важное влияние на механизмы функционирования белка.

Кроме того, отдельный исследовательский интерес вызывает группа пептидов нерибосомального пути биосинтеза благодаря их большому медицинскому значению. Среди них токсины, цитостатики, иммунодепрессоры [2]. В отличие от белков, кодируемых в геноме, синтез данной группы полипептидов происходит напрямую из аминокислот, что делает невозможным исследование первичной структуры методами геномики.

Это обусловливает исследование протеома клеток с помощью методов масс-спектроскопии. Тандемная масс-спектроскопия – это метод исследования, позволяющий установить первичную структуру пептидов [3]. В ее основе лежит (а) ферментативная фрагментация исследуемой смеси полипептидных цепей белков на короткие пептиды; (б) ионизация полученной смеси, при которой пептиды (их называют прекурсоры) получают заряд; (в) разделение этой пептидной смеси в ионной ловушке по отношению массы пептида m к его заряду z (m/z); (г) повторная диссоциация ионов-прекурсоров с образованием разнообразных структурно значимых ионных фрагментов, называемых вторичными ионами; (д) масс-анализ вторичных ионов. В результате такого разрушения исходной полипептидной цепи формируются серии спектров для каждого родительского иона. Каждый спектр – это распределение ионизированных фрагментов по величине отношения m/z и в графическом виде представляет собой набор узких пиков.

На следующем этапе осуществляется аннотация масс-спектров – сопоставление каждого из пиков масс-спектра пептида с известной структурной формулой. Аннотация позволяет определить пептидный состав белка. Информация о пептидном составе позволяет, в свою очередь, реконструировать последовательность исходного белка. Задача такой реконструкции является сложной, поскольку механизмы диссоциации пептидов до конца неизвестны, при аннотации установить структуру пептидов удается далеко не для всех спектральных пиков, на что отчасти влияет и совокупность внешних факторов, таких как величина заряда

прекурсора, величина энергии диссоциации и т. д. Дополнительные трудности создает посттрансляционная модификация белков при которой меняется химическая структура аминокислотных остатков уже после синтеза белка [4]. Посттрансляционная модификация существенно меняет значение *m*/*z* для аминокислот, кроме того, трудно заранее учесть все возможные варианты таких химических модификаций.

В настоящее время разработаны два типа алгоритмов реконструкции пептидной последовательности на основе масс-спектров. Первый основан на сопоставлении массы ионов в спектре с теоретическими масс-спектрами, рассчитанными на основании последовательностей всех возможных триптических пептидов для всех белков базы данных белковых последовательностей. Наличие уникальных теоретических пептидных фрагментов для какого-либо известного белка в спектре свидетельствует о его присутствии в смеси [3]. Ограничением таких алгоритмов является то, что в случае, если последовательность иона в базе отсутствует (что характерно для нерибосомальных пептидов), определение этого белка в смеси невозможно.

Алгоритмы сборки *de novo* работают на тех же принципах, что и алгоритмы сборки нуклеотидных последовательностей из фрагментов за счет их частичного перекрывания [1]. Они способны реконструировать последовательности пептидов, даже если те ранее были неизвестны. Однако процесс сборки требует большого времени счета (такая задача является NP-сложной [5]), и из-за того, что в спектрах всегда содержится шум (смещение линий из-за наличия изотопов, отсутствие некоторых линий, присутствие в спектре линий от загрязняющих веществ), получить решение весьма затруднительно. Также не ясна обобщающая способность алгоритмов *de novo* секвенирования для ранее неизвестных последовательностей.

Совершенствование существующих или разработка новых алгоритмов расшифровки масс-спектров требует много аннотированных масс-спектроскопических данных. Такие данные можно получить экспериментально, но этот способ обусловлен следующими недостатками: высокая трудоемкость, проведение экспериментов дорогостоящее, а аннотация масс-спектров часто проходит с невысокой достоверностью.

Другой способ получения таких данных – это компьютерное моделирование масс-спектров (масс-спектрометрия *in silico*). Данный подход не требует проведения трудоемких и дорогостоящих экспериментов, к тому же для масс-спектров оказываются известны исходные последовательности пептидов, что очень удобно для проверки работоспособности алгоритмов реконструкции.

В настоящее время разработан целый ряд методов, которые позволяют на основе аминокислотной последовательности пептида получить его спектр. Однако существующие алгоритмы моделирования масс-спектров не позволяют в полной мере проводить моделирование масс-спектров с учетом неканонического аминокислотного состава и посттрансляционных модификаций. Поэтому остается актуальной задача разработки алгоритмов моделирования масс-спектров пептидов.

В данной работе представлено построение алгоритма BioNet на основе методов искусственных нейронных сетей для моделирования масс-спектров пептидов с учетом неканонического аминокислотного состава и посттрансляционных модификаций. В работе проведено сравнение качества работы алгоритма с аналогами OpenMS-Simulator [6], MS2PIP [7] и pDeep [8].

Построение алгоритма BioNet

В настоящей работе мы предлагаем использовать для моделирования масс-спектров метод нейронных сетей. Нейронная сеть с точки зрения математики является вычислительным графом, в узлах которого содержатся операции или переменные, представляющие собой модели искусственных нейронов [9]. Графы искусственных нейронных сетей характеризуются наличием входного слоя, внутренних слоев и выходного слоя. Искусственные нейронные сети,

заданные таким образом, способны приблизить любую непрерывную функцию с любой требуемой точностью, в том числе и для задач обработки естественного языка – более общего случая обработки последовательностей.

В нашей работе, чтобы с помощью нейронной сети смоделировать по заданной аминокислотной последовательности масс-спектр, последовательность пептида необходимо представить в векторном виде, в котором ее можно будет поместить на входной слой вычислительного графа. Чтобы сохранять информацию о порядке аминокислот, удобно представлять последовательность пептида в виде последовательности признаков, которые уже кодируют отдельные аминокислоты. Использование признаков на основе векторов скрытого представления позволяет избежать недостатков вышеперечисленных подходов. В данной работе использовались векторы скрытого представления, полученные с помощью модели Mol2Vec [10].

Модель Mol2Vec в качестве предложения использует химическую структуру молекулы пептида, а для построения «слов» на основе этой структуры используется алгоритм Моргана [11], который для каждого атома производит идентификатор (рис. 1). Из этих идентификаторов составляется «предложение», которое характеризует молекулу. На вход модели Mol2vec подается химическая структура молекулы в формате SMILES [12].

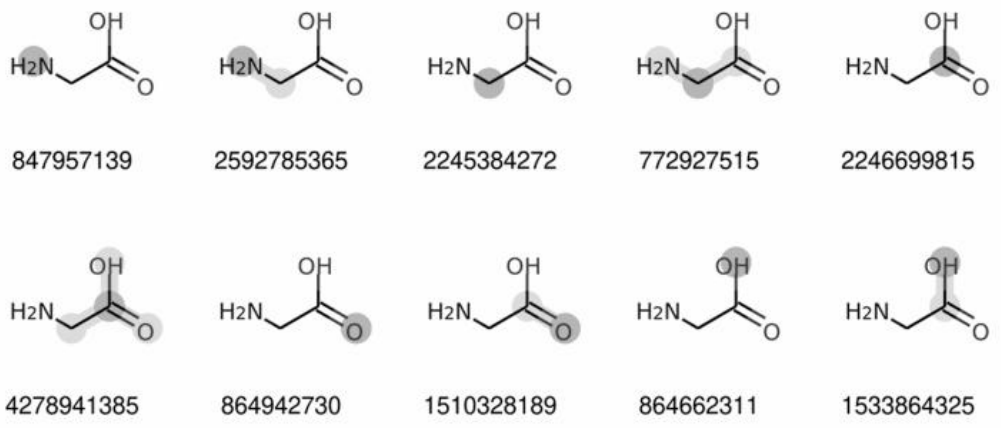

Рис. 1. Пример построения «слов» с помощью алгоритма Моргана для глицина. Идентификаторы расположены в том же порядке, что и атомы для канонического представления SMILES. Если атом имеет более чем один идентификатор, то они расположены в порядке возрастания радиуса (расстояние между атомами в графе молекулы). Адаптировано из [12]

Fig. 1. The example of construction “words” with Morgan algorithm for glycine. Identifiers is placed the same order as atoms in SMILES canonical representation of molecule. If atom has more than one identifier, then they are placed in ascending order. Adapted from [12]

В данной работе использована модель Mol2Vec типа Skipgram с размером контекста 10, радиусом Моргана 1 и размерностью векторов скрытого пространства 300. Модель доступна по адресу github.com/samoturk/mol2vec.

После определения способа векторизации последовательностей пептидов дальнейший поиск архитектуры необходимо было проводить в терминах задачи моделирования последовательности по последовательности. В контексте решаемой задачи требуется по аминокислотной последовательности пептида моделировать интенсивности ионов основных серий, получаемых в ходе масс-спектроскопического эксперимента. В таких задачах моделирования хорошо себя зарекомендовали архитектуры нейронных сетей на рекуррентных моделях [13; 14].

Поскольку на величину энергии разрыва связи в пептиде влияют аминокислоты как со стороны N-конца, так и со стороны C-конца [8], принято решение использовать двунаправленный вариант LSTM слоев, чтобы учесть данную особенность. Кроме того, на величину энергии также оказывают влияние концевые аминокислоты, поэтому логично передавать на вход нейронной сети не только последовательность пептида, но и бинарную маску, в которой будут помечены концевые аминокислоты последовательности.

В [15] показано, что обучаемая линейная комбинация выходов слоев дает лучшее представление для решения задачи. Такая же техника использована в данной работе. Итоговая архитектура нейронной сети для моделирования масс-спектров представлена на рис. 2.

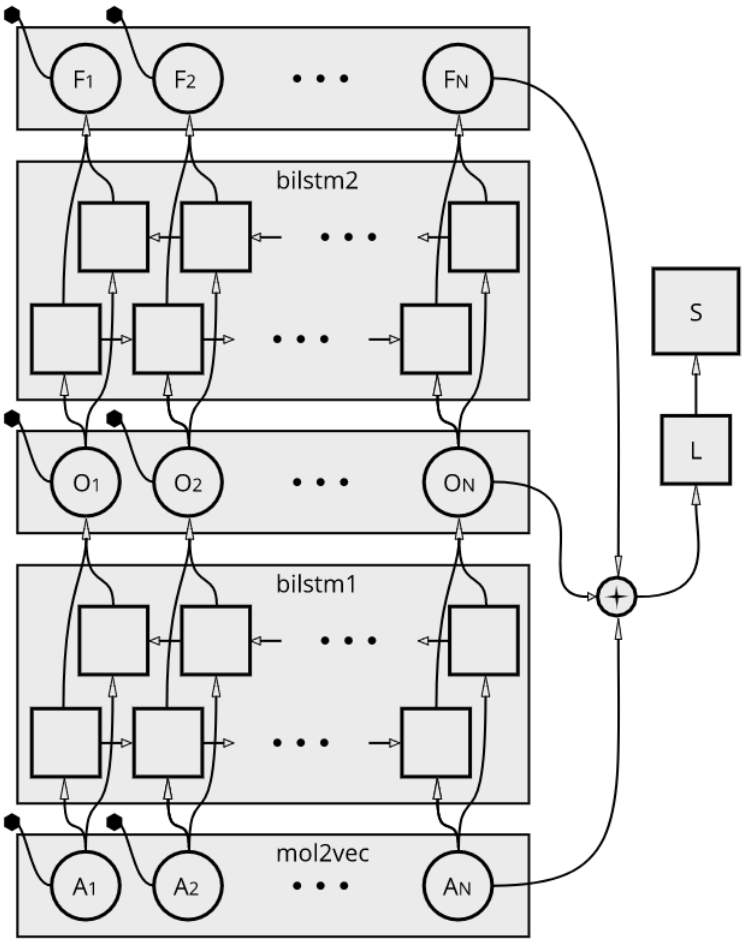

Рис. 2. Архитектура нейронной сети для моделирования масс-спектров пептидов: Ai – аминокислота; Oi – выход первого двунаправленного слоя; Fi – выход второго двунаправленного слоя; mol2vec – слой сети для получения векторов скрытого представления для аминокислот; bilstm – двунаправленный рекуррентный слой; + – обучаемая линейная комбинация, • – действие обучаемой линейной комбинации; L – группа линейный слоев; S – смоделированный спектр

Fig. 2. Neural network architecture to predict peptide mass-spectra. Ai – amino acid, Oi – first bidirectional layer output, Fi – second bidirectional layer output, mol2vec – amino acid embedding layer, bilstm – bidirectional recurrent layer, + – learnable linear combination, • – action learnable linear combination, L – stacked linear layer, S – predicted mass-spectrum

Описание метрики качества и функции потерь для обучения алгоритма

В задаче попарного сравнения масс-спектров принято использовать следующие метрики: косинусное расстояние [8; 16; 17], коэффициенты корреляции Пирсона [7; 8] или Спирмена [8]. В данной работе используется коэффициент корреляции Пирсона (ККП) для попарного сравнения масс-спектров. ККП характеризует существование линейной зависимости между двумя величинами.

Пусть дана аминокислотная последовательность, известен экспериментальный масс-спектр y и смоделирован теоретический масс-спектр p.

ККП рассчитывается по формуле

$$pearsim\ \ y, p\ \ = \frac{\sum_j \left(y_j - \underline{y}\right)\left(p_j - \underline{p}\right)}{\left[\sum_j^2 \left(y_j - \underline{y}\right)\right]^{0.5}\left[\sum_j^2 \left(p_j - \underline{p}\right)\right]^{0.5}},$$

где $\underline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$.

Оценки качества моделирования масс-спектров для набора D пар масс-спектр-пептид происходит следующим образом:

$$pD = pearson(D),$$
$$pD = sort(pD),$$
$$mpD = median(pD),$$
$$rpD = \frac{1}{|pD|}\sum_{i\ pD_i > 0.75},$$

где mpD – медианное значение (мККП), а rpD – доля больше 0,75 (дККП) для набора ККП. Величины мККП и дККП характеризуют качество моделирования масс-спектров: чем они больше, тем выше точность моделирования масс-спектров.

Для обучения сети использована функция потерь следующего вида:

$$loss = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - p_i\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_j \left(y_{ij} - \underline{y_i}\right)\left(p_{ij} - \underline{p_i}\right)}{\left[\sum_j^2 \left(y_{ij} - \underline{y_i}\right)\right]^{0.5}\left[\sum_j^2 \left(p_{ij} - \underline{p_i}\right)\right]^{0.5}} + \sum_k \frac{c}{w_k^2},$$

где y_i – экспериментальный масс-спектр, p_i – теоретический масс-спектр, w_k – веса нейронной сети, C – константа регуляризации, N – размер обучающей выборки. Первый член функции потерь оптимизирует абсолютные значения смоделированного спектра, второй член оптимизирует метрику качества, последний член необходим для регуляризации весов модели.

Выбор данных для обучения алгоритма

В качестве источника данных для тестирования алгоритма BioNet использовались масс-спектры из базы данных PRIDE [18], проекты PXD004732 и PXD000138. Проект PXD004732 содержит около четырех миллионов аннотированных тандемных масс-спектров пептидов с каноническим аминокислотным составом [19]. Этот набор данных получен в рамках проек-

та Proteome Tools. Масс-спектры Proteome Tools получены в результате расщепления пептидов человека трипсином, дальнейшей диссоциации на разных режимах с последующим использованием различных вариантов детекторов ионов. Полученные масс-спектры аннотировались с помощью программы Andromeda [20], которая является поисковой системой, использующей информацию о достоверно определенных масс-спектрах пептидов для аннотации неизвестных масс-спектров.

Проект PXD000138 содержит около двухсот тысяч масс-спектров, полученных методом диссоциации HCD, для пептидов с каноническим аминокислотным составом и их аналогов с модификациями типа оксидирование (M), фосфорилирование (ST) и фосфорилирование (Y) [21]. Масс-спектры в рамках данного проекта также аннотировались с помощью программы Andromeda [20].

Обработка данных

Для загруженных данных были построены распределения пептидов по длине и заряду (рис. 3, 4), для того чтобы оценить эти две основные характеристики пептидных данных.

На основе построенных распределений было решено ограничить размер входной последовательности двадцатью аминокислотами и исключить из рассмотрения масс-спектры зарядом прекурсора один, пять и более из-за их малой представленности в данных. Ранее в работе Zolg и др. [22] было предложено использовать только результаты аннотации с коэффициентом уверенности (Andromeda score) больше 100. Мы использовали этот порог для фильтрации наших данных. Кроме того, масс-спектр исключался из рассмотрения, если количество аннотированных пиков было меньше, чем количество аминокислот в пептиде.

После фильтрации данных по количеству аминокислот в пептиде, заряду прекурсора, качеству аннотации и количеству пиков из проекта PXD004732 было получено 2 325 000 масс-спектров, из проекта PXD000138 – 91 230 масс-спектров, среди них 36 740 масс-спектров с модифицированными аминокислотными последовательностями.

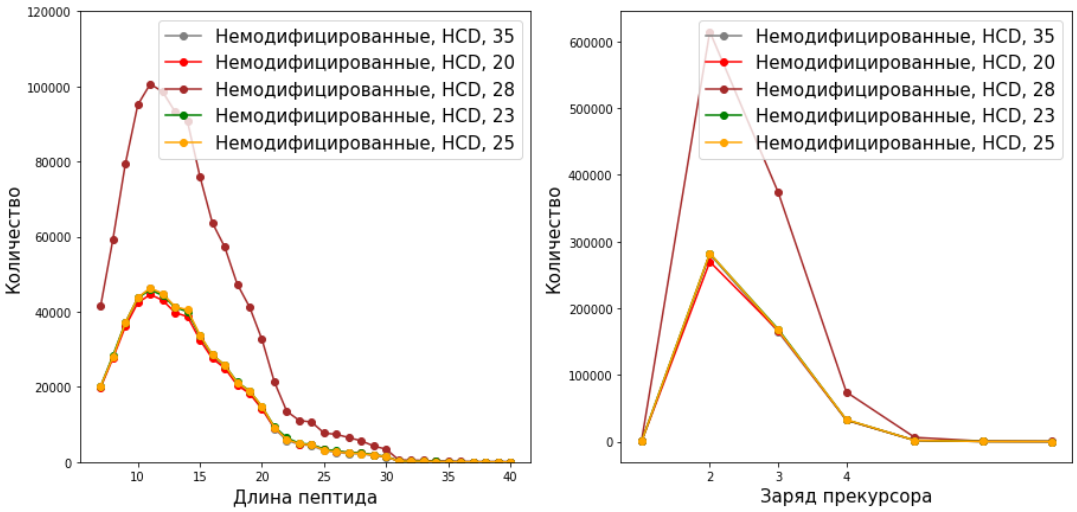

Рис. 3. График распределения пептидов по длине (слева) и заряду прекурсора (справа) для данных проекта PXD004732

Fig. 3. Peptide distribution plots by length (left) and precursor charge (right) over PXD004732 data

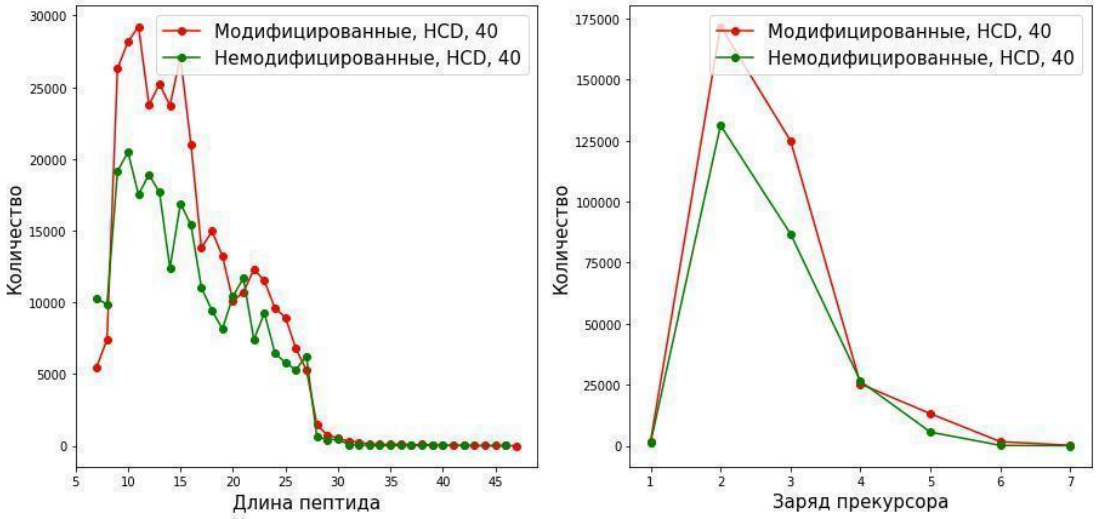

Рис. 4. График распределения пептидов по длине (слева) и заряду прекурсора (справа) для данных проекта PXD000138

Fig. 4. Peptide distribution plots by length (left) and precursor charge (right) over PXD000138 data

Для обучения было отобрано 80 % данных с каноническим аминокислотным составом, остальные 20 % были отложены для тестирования, также 100 % данных с неканоническим аминокислотным составом были использованы для тестирования.

Точность моделирования масс-спектров

Результатом работы нашего алгоритма является смоделированный масс-спектр пептида, поэтому эффективность работы алгоритма формулируется в терминах мКПП и дКПП, описанных ранее. Чем выше эти значения, тем точнее работает алгоритм.

мККП и дККП, полученные в результате тестирования разработанного метода, приведены в табл. 1, 2. Данные представлены в сравнении с алгоритмами OpenMS-Simulator, MS2PIP и pDeep. Также для данных была получена теоретическая оценка максимально достижимого качества [8], расчет проводился для последовательностей, представленных несколькими масс-спектрами, моделированным масс-спектром считался усредненный масс-спектр. Пример смоделированных масс-спектров с ККП 0.992 и 0.718 приведен на рис. 5.

Таблица 1

Результаты тестирования для PDX000138, PDX004732, канонические аминокислоты

Table 1

The results of evaluation over PDX000138, PDX004732 data with canonical amino acid composition

	мККП	дККП
OpenMS-Simulator	0.710	44.05
MS2PIP	0.845	71.80
pDeep	0.246	4.36
Предложенный алгоритм	0.909	76.78
Верхняя оценка	0.986	99.33

Таблица 2

Результаты тестирования для PDX000138, посттрансляционные модификации (фосфорилирование по S, T, Y, оксидирование по M) *

Table 2

The results of evaluation over PDX000138 data with posttranslational modifications (phospho S, T, Y, oxidation M) **

	мККП	дККП
OpenMS-Simulator	0.710	44.05
MS2PIP	0.845	71.80
pDeep	0.246	4.36
Предложенный алгоритм	0.909	76.78
Верхняя оценка	0.986	99.33

* Приведены данные моделирования без учета модификаций.
** Data are predicted disregarding posttranslational modifications.

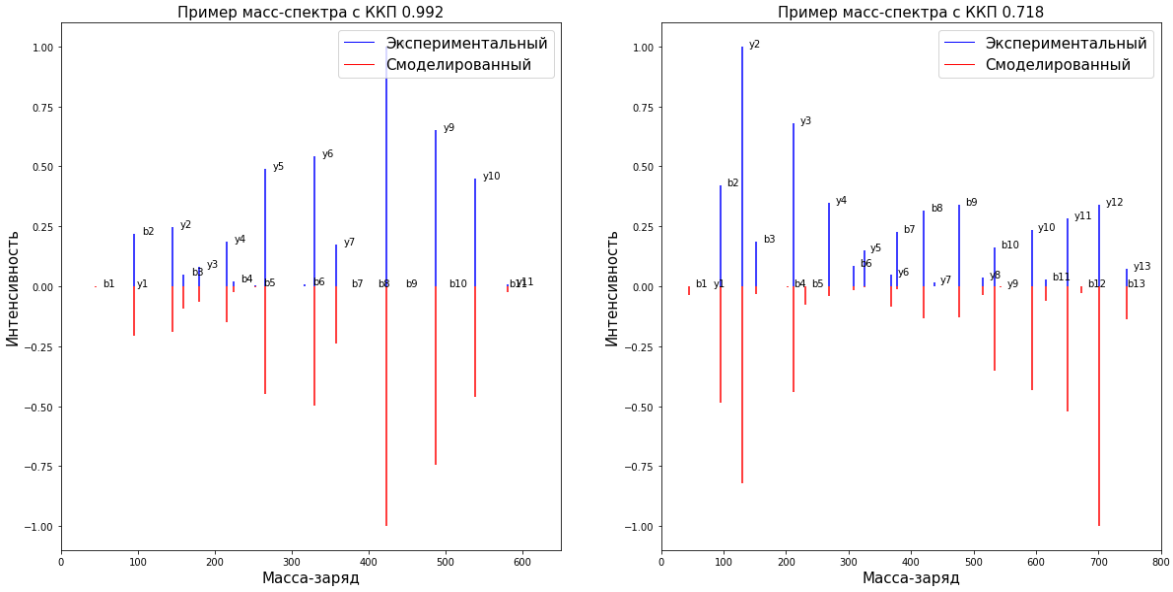

Рис. 5. Пример моделирования масс-спектра предложенным алгоритмом. Сверху расположен экспериментальный масс-спектр, снизу теоретический. Спектры представлены для последовательностей STEEGEVAALR (слева) и STDVGRHSLLYLK (справа)

Fig. 5. The example of predicting mass-spectrum by developed algorithm. From above it is placed experimental, from below is placed theoretical mass-spectrum. Spectra are shown for STEEGEVAALR (left) and STDVGRHSLLYLK (right) sequences

Заключение

В рамках выполнения работы были получены следующие результаты.

1. Реализован алгоритм BioNet для моделирования масс-спектров пептидов на основе нейронной сети глубокого обучения.

2. Проведено сравнение разработанного метода и трех аналогичных алгоритмов для задачи моделирования масс-спектров пептидов с каноническим аминокислотным составом. Разработанный алгоритм показал лучшее качество моделирования.

3. Проведено сравнение разработанного метода для моделирования масс-спектров с неканоническим аминокислотным составом. Разработанный алгоритм показал наиболее высокое качество моделирования спектров.

Список литературы

1. **Задесенец К. С., Ершов Н. И., Рубцов Н. Б.** Полногеномное секвенирование геномов эукариот: от секвенирования фрагментов ДНК к сборке генома // Генетика. 2017. Т. 53, № 6. С. 641–650.
2. **Орлова Т. И., Булгакова В. Г., А. Н. Полин.** Биологически активные нерибосомальные пептиды. III. Нерибосомальные антибиотики полипептиды // Антибиотики и химиотерапия. 2012. № 7. С. 43–54.
3. **Краснов Н. В., Лютвинский Я. И., Подольская Е. П.** Масс-спектрометрия с мягкими методами ионизации в протеомном анализе (Обзор) // Научное приборостроение. 2010. Т. 20, № 4. С. 5–20.
4. **Кнорре Д. Г., Кудряшова Н. В., Годовикова Т. С.** Химические и функциональные аспекты посттрансляционной модификации белков // Acta Naturae. 2009. Т. 1, № 3.
5. **Fomin E. A.** Simple Approach to the Reconstruction of a Set of Points from the Multiset of n 2 Pairwise Distances in n 2 Steps for the Sequencing Problem: I. Theory. *Journal of Computational Biology*, 2016, vol. 23, no. 9, p. 769–775.
6. **Wang Y. et al.** OpenMS-Simulator: an open-source software for theoretical tandem mass spectrum prediction. *BMC bioinformatics*, 2015, vol. 16, no. 1, p. 110.
7. **Degroeve S., Martens L.** MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, 2013, vol. 29, no. 24, p. 3199–3203.
8. **Zhou X. X. et al.** pDeep: predicting MS/MS spectra of peptides with deep learning. *Analytical chemistry*, 2017, vol. 89, no. 23, p. 12690–12697.
9. **Созыкин А. В.** Обзор методов обучения глубоких нейронных сетей // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2017. Т. 6, № 3. С. 28–59. DOI 10.14529/cmse170303
10. **Jaeger S., Fulle S., Turk S.** Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 2018, vol. 58, no. 1, p. 27–35.
11. **Rogers D., Hahn M.** Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 2010, vol. 50, no. 5, p. 742–754.
12. **Richard L. Rowley, R. Jeremy Rowley, John L. Oscarson, W. Vincent Wilding.** Development of an Automated SMILES Pattern Matching Program to Facilitate the Prediction of Thermo physical Properties by Group Contribution Methods. Department of Chemical Engineering. Brigham Young University. Provo, Utah, 2001, p. 1110–1113.
13. **Sutskever I., Vinyals O., Le Q. V.** Sequence to Sequence Learning with Neural Networks. In: arXiv preprint arXiv:1409.3215. 2014.
14. **Graves A.** Generating sequences with recurrent neural networks. In: arXiv preprint arXiv:1308.0850. 2013.
15. **Peters M. E. et al.** Deep contextualized word representations. In: arXiv preprint arXiv:1802.05365. 2018.
16. **Dong N. et al.** Prediction of peptide fragment ion mass spectra by data mining techniques. *Analytical chemistry*, 2014, vol. 86, no. 15, p. 7446–7454.
17. **Lin Y. M., Chen C. T., Chang J. M.** MS2CNN: predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks. *BMC genomics*, 2019, vol. 20, no. 9, p. 1–10.
18. **Perez-Riverol Y. et al.** The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research*, 2019, vol. 47, no. D1, p. D442–D450.

19. **Gessulat S. et al.** Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 2019, vol. 16, no. 6, p. 509.
20. **Cox J. et al.** Andromeda: a peptide search engine integrated into the MaxQuant environment. **Journal of proteome research**, 2011, vol. 10, no. 4, p. 1794–1805.
21. **Marx H. et al.** A large synthetic peptide and phosphopeptide reference library for mass spectrometry based proteomics. *Nature biotechnology*, 2013, vol. 31, no. 6, p. 557–564.
22. **Zolg D. P. et al.** Building Proteome Tools based on a complete synthetic human proteome. *Nature methods*, 2017, vol. 14, no. 3, p. 259–262.

References

1. **Zadesenets K. S., Yershov N. I., Rubtsov N. B.** Genome-Wide sequencing of genomes eukaryote: from sequencing of DNA fragments for Assembly of the genome. *Genetics*, 2017, vol. 53, no. 6, p. 641–650. (in Russ.)
2. **Orlova T. I., Bulgakova V. G., Polin A. N.** Biologically active non-ribosomal peptides. III. Non-ribosomal antibiotics polypeptides. *Antibiotics and Chemotherapy*, 2012, no. 7, p. 43–54. (in Russ.)
3. **Krasnov N. V., Lyutvinsky Ya. I., Podolskaya E. P.** Soft mass spectrometry methods of ionization in proteomic analysis (Review). Scientific Instrumentation, 2010, vol. 20, no. 4, p. 5–20. (in Russ.)
4. **Knorre D. G., Kudryashova N. V., Godovikova T. S.** Chemical and functional aspects of posttranslational modification of proteins. *Acta Naturae*, 2009, vol. 1, no. 3. (in Russ.)
5. **Fomin E. A.** Simple Approach to the Reconstruction of a Set of Points from the Multiset of n 2 Pairwise Distances in n 2 Steps for the Sequencing Problem: I. Theory. *Journal of Computational Biology*, 2016, vol. 23, no. 9, p. 769–775.
6. **Wang Y. et al.** OpenMS-Simulator: an open-source software for theoretical tandem mass spectrum prediction. *BMC bioinformatics*, 2015, vol. 16, no. 1, p. 110.
7. **Degroeve S., Martens L.** MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, 2013, vol. 29, no. 24, p. 3199–3203.
8. **Zhou X. X. et al.** pDeep: predicting MS/MS spectra of peptides with deep learning. *Analytical chemistry*, 2017, vol. 89, no. 23, p. 12690–12697.
9. **Sozykin A. V.** Review of training methods for deep neural networks. *Bulletin of SUSU. Series: Computational mathematics and computer science*, 2017, vol. 6, no. 3, p. 28–59. (in Russ.) DOI 10.14529/cmse170303
10. **Jaeger S., Fulle S., Turk S.** Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 2018, vol. 58, no. 1, p. 27–35.
11. **Rogers D., Hahn M.** Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 2010, vol. 50, no. 5, p. 742–754.
12. **Richard L. Rowley, R. Jeremy Rowley, John L. Oscarson, W. Vincent Wilding.** Development of an Automated SMILES Pattern Matching Program to Facilitate the Prediction of Thermo physical Properties by Group Contribution Methods. Department of Chemical Engineering. Brigham Young University. Provo, Utah, 2001, p. 1110–1113.
13. **Sutskever I., Vinyals O., Le Q. V.** Sequence to Sequence Learning with Neural Networks. In: arXiv preprint arXiv:1409.3215. 2014.
14. **Graves A.** Generating sequences with recurrent neural networks. In: arXiv preprint arXiv:1308.0850. 2013.
15. **Peters M. E. et al.** Deep contextualized word representations. In: arXiv preprint arXiv:1802.05365. 2018.
16. **Dong N. et al.** Prediction of peptide fragment ion mass spectra by data mining techniques. *Analytical chemistry*, 2014, vol. 86, no. 15, p. 7446–7454.

17. **Lin Y. M., Chen C. T., Chang J. M.** MS2CNN: predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks. *BMC genomics*, 2019, vol. 20, no. 9, p. 1–10.
18. **Perez-Riverol Y. et al.** The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research*, 2019, vol. 47, no. D1, p. D442–D450.
19. **Gessulat S. et al.** Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 2019, vol. 16, no. 6, p. 509.
20. **Cox J. et al.** Andromeda: a peptide search engine integrated into the MaxQuant environment. **Journal of proteome research**, 2011, vol. 10, no. 4, p. 1794–1805.
21. **Marx H. et al.** A large synthetic peptide and phosphopeptide reference library for mass spectrometry based proteomics. *Nature biotechnology*, 2013, vol. 31, no. 6, p. 557–564.
22. **Zolg D. P. et al.** Building Proteome Tools based on a complete synthetic human proteome. *Nature methods*, 2017, vol. 14, no. 3, p. 259–262.

Материал поступил в редколлегию
Received
05.06.2020

Сведения об авторах

Епифанов Ростислав Юрьевич, магистрант 2-го курса факультета информационных технологий Новосибирского национального исследовательского государственного университета (Новосибирск, Россия)
rostepifanov@gmail.com

Афонников Дмитрий Аркадьевич, кандидат биологических наук, ведущий научный сотрудник Института цитологии и генетики СО РАН (Новосибирск, Россия)
ada@bionet.nsc.ru

Information about the Authors

Rostislav Yu. Epifanov, Master Student 2yr, Faculty of Information Technologies, Novosibirsk State University (Novosibirsk, Russian Federation)
rostepifanov@gmail.com

Dmitry A. Afonnikov, PhD of Biology, Leading Research Scientist, ICG SB RAS (Novosibirsk, Russian Federation)
ada@bionet.nsc.ru