Обнаружение и анализ белковых соединений на основе рамановского рассеяния и машинного обучения

А.С. Штумпф*
Национальный исследовательский университет ИТМО
Россия, 197101, Санкт-Петербург, Кронверкский пр., д. 49
(Поступила в редакцию 09.06.2024; подписана в печать 05.07.2024)

Белки являются важными компонентами человеческого организма, выполняя решающую роль в функционировании клеток: катализируют химические реакции и формируют клеточные структуры. Дисбаланс белкового обмена может иметь серьезные последствия, такие как нарушение иммунитета и изменение активности желез. Обнаружение различных биологических соединений представляет собой проблему из-за их сложных межмолекулярных связей, а традиционные методы, такие как иммуноанализ и хроматография, не всегда могут дать точные результаты. Представленное исследование направлено на преодоление этих ограничений путем внедрения подхода, который объединяет рамановскую спектроскопию и машинное обучение для точной идентификации белковых соединений. Эта методика направлена на минимизацию ошибок в количественном и качественном анализе и обеспечение систематического исследования белковых соединений. Полученные в ходе тестирования алгоритма на полученных в ходе экспериментов данных результаты свидетельствуют о возможности применения такой методики для более чем 10 веществ-аналитов и достижении точности свыше 90%. Сформированная таким образом методика работы с экспериментальными данными с использованием средств искусственного интеллекта может лечь в основу создания эффективных платформ и устройств для применения не только в научной сфере, но и также в сферах медицины, сельского хозяйства, пищевой безопасности.

PACS: 42.65.Dr. УДК: 53.06.
Ключевые слова: аминокислоты, белки, гормоны, рамановская спектроскопия, машинное обучение, ранняя диагностика заболеваний.

ВВЕДЕНИЕ

Обнаружение различных биологических соединений является трудоемким процессом из-за сложности их межмолекулярных связей. Современные методы иммуноанализа и хроматографии не всегда позволяют добиться результатов в короткие сроки и с использованием небольшого количества ресурсов [1, 2]. Таким образом, научная задача, которую призвана решить данная работа, — это разработка быстрого и простого в использовании метода, позволяющего добиться хороших результатов для задач, связанных с обнаружением сложных биологических соединений, в частности для распознавания гормонов [3–6].

Используемый в работе подход включает анализ спектров комбинационного рассеяния аминокислот и более сложных белковых соединений, а также применение алгоритмов машинного обучения для прогнозирования значений концентраций и идентификации компонентов смеси [7].

1. МАТЕРИАЛЫ И МЕТОДЫ

Получены рамановские спектры высушенных водных 0.1 М растворов двух одиночных аминокислот (аланина и глутамина), нанесенных на стекло. В ходе измерений использовался рамановский спектрометр (Renishaw, Англия), оснащенный HeNe-лазером с длиной волны 633 нм и максимальной мощностью 5 мВт. Помимо сигналов от чистых аминокислот, были получены сигналы от смесей этих двух аминокислот в разных процентных соотношениях. Спектры смеси с 50% концентрацией как аланина, так и глутамина были получены для сравнения этих сигналов со сигналами дипептида L-аланил-L-глутамин в концентрации 0.1 М.

Обработка спектров, полученных в ходе измерений, состоящая из нахождения и вычитания базовой линии, нормализации данных, сглаживания спектров по алгоритму Савицкого–Голея, поиска значений частот, соответствующих пикам интенсивности, осуществлялся с использованием библиотек языка программирования Python.

Проведено сравнение рамановского сигнала смеси двух аминокислот с ожидаемыми спектрами, полученными сложением сигналов отдельных аминокислот с заданными коэффициентами. Спектры комбинационного рассеяния смесей аминокислот обрабатывались с помощью алгоритма, позволяющего путем нахождения минимума функции ошибок теоретически построенного спектра методом Лагранжа определять значение концентраций компонентов смеси. Квадрат функции ошибок определяется следующим уравнением:

$$f(x) = \sum_{i=1}^{1011} \left(y_i - (x_{i0}\theta_0 + \ldots + x_{in}\theta_n)\right)^2, \qquad (1)$$

* artem.shtumpf@metalab.ifmo

где θ_k — компоненты вектора, содержащего значения концентраций каждой из аминокислот, входящей в заданную смесь; y_i — значение интенсивности излучения для конкретного значения длины волны в спектре рамановского рассеяния смеси в выбранном спектральном диапазоне; x_{ik} — значение интенсивности при заданном значении длины волны в спектре одиночной аминокислоты (индекс «i» соответствует точки спектра, индекс «k» соответствует компоненте вектора концентраций θ_k).

Затем были использованы алгоритмы машинного обучения для оптимизации параметров системы (с использованием метода главных компонент), прогнозирования значений параметров на основе известных данных (с использованием логистической регрессии) и точной идентификации компонент в смесях на основе спектров комбинационного рассеяния света (с использованием деревьев решений, случайных лесов и машины опорных векторов).

2. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

2.1. Смесь из двух аминокислот. Идентификация

На первом этапе работы были рассмотрены две протеиногенные аминокислоты: аланин, глутамин, для которых получены сигналы рамановского рассеяния. Были получены сигналы рамановского рассеяния от смеси двух аминокислот в разном процентном соотношении (100:0, 80:20, 75:25, 60:40, 50:50).

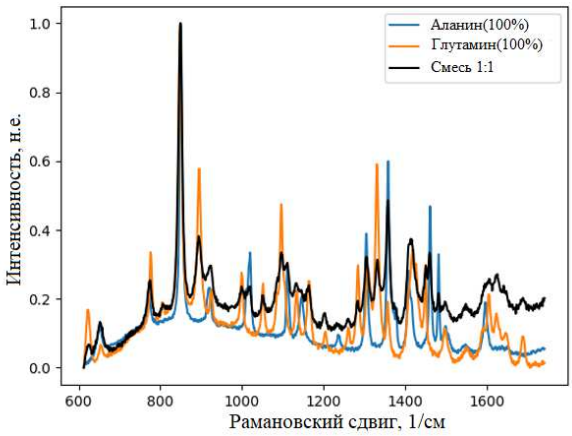

Рис. 1. Усредненные и предварительно обработанные спектры комбинационного рассеяния чистых аминокислот: аланина (100%) и глутамина (100%), где 100% соответствует концентрации конкретной аминокислоты в растворе; смесь аланина и глутамина в равных пропорциях в растворе

Для извлечения ключевых особенностей спектральных данных для рассматриваемого класса веществ были применены различные техники предобработки данных. Основной целью такой предобработки является исключение определенных вариаций спектра, не несущих в себе определяющую для выбранного метода анализа информацию. В контексте применения методов рамановской спектроскопии важным является исключения из спектральных данных информации о проявляющейся при экспериментах люминесценции, инструментальной погрешности, а также шумах. Для решения данной задачи применялись техники предобработки спектральных данных: вычет базовой линии, метод Multiplicative Scatter Correction(MSC) для уменьшения разброса в данных, алгоритм Савицкого–Голея для избавления от шумов.

Для применения алгоритмов искусственного интеллекта путём обучения модели было необходимо перевести исходные спектральные данные в определенное представление (рис. 2). Использовалось несколько подходов:

1. сопоставление частот из выбранного диапазона частот (600–1800 $см^{-1}$) значениям интенсивности в этих точках;

2. определение частот, соответствующих пикам интенсивности заданного спектра и сопоставление этих точек значениям интенсивности;

3. поиск пиков интенсивности и суммирование значений интенсивности по нескольким точкам в окрестности пика.

Причина применения различных подходов заключается в необходимости контролировать качество обучения модели искусственного интеллекта. Это может быть сделано путем построения и анализа так называемой кривой обучения, представляющей из себя зависимость точности предсказания модели от количества используемых данных [8]. В процессе подготовки модели и определения соответствующих гиперпараметров возможно проявление проблем недообучения или переобучения модели [9, 10]. В связи с возможностью возникновения таких проблем менялся способ представления спектральных данных, прямым образом влияющий на количество предикторов — признаков, на основании значений которых строится предсказание модели.

Методы машинного обучения использовались для классификации различных соединений на основе их спектров комбинационного рассеяния света. Методы включали двухклассовую классификацию для сравнения спектров чистого аланина и глутамина, а также трехклассовую классификацию для анализа аланина, глутамина и их смесей. Алгоритмы KNN и Random Forest были применены и оказались эффективными для этих задач. Перекрестная проверка использовалась для оптимизации параметров модели, в результате чего точность и коэффициент полноты превысили 0,96 для набора обучающих данных из 2000 спектров (таблица).

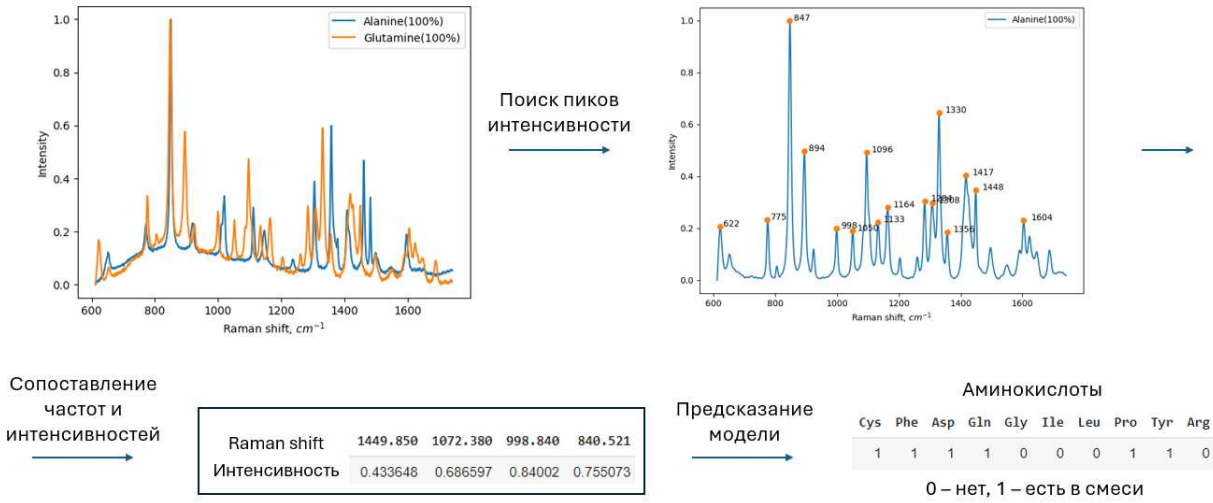

Рис. 2. Схема перевода спектральных данных в представление для обучения алгоритма искусственного интеллекта

Таблица. Результаты двухклассовой классификации

	Precision	Recall	f1-score	Количество спектров в тестовом наборе данных
Аланин	0.97	0.99	0.98	73
Глутамин	0.86	1.00	0.92	72
Смесь	1.00	0.94	0.96	218
Accuracy			0.96	363
Macro AVG	0.94	0.97	0.96	363
Weighed AVG	0.96	0.96	0.96	363

3. Результаты двухклассовой классификации

2.2. Смесь из двух аминокислот. Количественная характеризация

Задача количественной характеризации образцов в рамках работы сводилась к определению концентрацией компонент смеси по рамановским спектрам. Сначала был рассмотрен упрощенный случай, когда спектр итоговой смеси был представлен как суперпозиция спектров отдельных ел компонент с линейными коэффициентами. Фактически такой же подход лежит в основе метода Линейной регрессии, в котором происходит вычисление суммы входных признаков в выбранном наборе данных со своими весами, оптимизация которых проводится в процессе минимизации функции ошибок. Аналогичным образом происходило выполнение предсказания моделью в методе Логистической Регрессии. Однако в том случае вычислялось не само значение выбранной метки, а вероятность отнесения задаваемого в нашем случае спектром объекта к определенному классу.

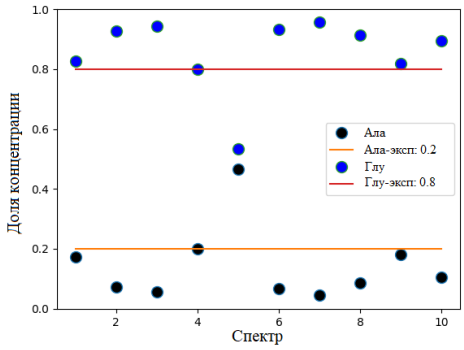

Рис. 3. Результаты концентрационного расчета для нескольких полученных спектров аланин-глутаминовой смеси: 20% аланина(Ала), 80% глутамина(Глу). Значения рассчитанных концентраций показаны синими и черными кружками; ожидаемые спектры, полученные в результате объединения сигналов отдельных аминокислот с указанными коэффициентами, показаны горизонтальными линиями (Ала-эксп, Глу-эксп)

Определение коэффициентов происходило путем минимизации функции ошибок, связывающей реальный спектр смеси и такую линейную суперпозицию, с помощью метода Лагранжа (рис. 3).

ЗАКЛЮЧЕНИЕ

L Результаты применения методов машинного обучения к полученным спектрам рамановского рассеяния свидетельствуют о возможности получения точных результатов при решении задачи классификации для простейших систем, состоящих из различных аминокислот. Это позволяет перейти к анализу более сложных систем, составными частями которых являются три и более аминокислот. В дальнейшем планируется провести такой анализ с применением более сложных моделей искусственного интеллекта, в том числе ансамблевых моделей (бэггинг, бустинг). Этот подход может лечь в основу анализа более сложных, в первую очередь, по своей структуре молекул различных белковых соединений, например, гормонов. Сформированная и отработанная таким образом методика работы с экспериментальными данными может лечь в основу создания эффективных платформ и устройств для применения не только в научной сфере, но и также в сферах медицины, сельского хозяйства, пищевой безопасности.

Исследования выполнены при поддержке РНФ (грант № 21-72-30018, ссылка на информацию о проекте: https://rscf.ru/project/21-72-30018/).

[1] *Belachew B. Hirpessa, Beyza H. Ulusoy, Canan Hecer.* // J. Food Qual. Aug. (2020).
[2] *Максимова Н.Е., Мочульская Н.Н., Емельянов В.В.* / Основы иммуноанализа: учебное пособие. http://hdl.handle.net/10995/106083(2021).
[3] *Francjan J. van Spronsen.* et al // Nat. Rev. Dis. Primers **7**(36), 1. May (2021).
[4] Альманах классической медицины. **48**(4) 254 (2020).
[5] *Mirjam Christ-Crain* et al. // In: Nat. Rev. Dis. Primers 5.54, pp. 1–20(Aug. 2019).
[6] *Kleindienst A., Hannon M.J., Buchfelder M., Verbalis J.G.* // Journal of Neurotrauma. **33**(7). 615 (2016).
[7] *Hatice Altug* et al. // Nat. Nanotechnol. **17**. 5 (Jan. 2022).
[8] *Anzanello M.J., Fogliatto* F.S. // Int. J. Ind. Ergon. **41**(5). 573 (Sept. 2011).
[9] *Goodfellow I., Bengio Y., Courville A.* Deep Learning (Adaptive Computation and Machine Learning series). Cambridge, MA, USA: The MIT Press, Nov. 2016.ISBN: 978-0-26203561-3. https://www.amazon.com/DeepLearningAdaptiveComputationMachine/dp/0262035618
[10] *Gareth J.* et al. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics). New York, NY, USA: Springer, June 2013. ISBN:978-1-46147137-0. https://www.amazon.com/Introduction-Statistical-Learning--Applications-Statistics/dp/1461471370

Identification and examination of protein compounds utilizing Raman scattering and machine learning techniques

A.S. Shtumpf

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University)
Saint-Petersburg 197101, Russia
E-mail: artem.shtumpf@metalab.ifmo.ru

Proteins are important components of the human body, playing a crucial role in the functioning of cells: they catalyze chemical reactions and form cellular structures. An imbalance in protein metabolism can have serious consequences, such as impaired immunity and changes in glandular activity. Detection of various biological compounds is challenging due to their complex intermolecular relationships, and traditional methods such as immunoassays and chromatography may not always provide accurate results. The presented research aims to overcome these limitations by introducing an approach that combines Raman spectroscopy and machine learning to accurately identify protein compounds. This technique aims to minimize errors in quantitative and qualitative analysis and enable systematic investigation of protein compounds.The results obtained during testing of the algorithm on data obtained during experiments indicate the possibility of using this technique for more than 10 analyte substances and achieving an accuracy of over 90%. The methodology for working with experimental data using artificial intelligence tools thus formed can form the basis for creating effective platforms and devices for use not only in the scientific field, but also in the fields of medicine, agriculture, and food safety.

PACS: 42.65.Dr.
Keywords: amino acids, proteins, hormones, raman spectroscopy, machine learning, early diagnosis of diseases.
Received 09 June 2024.

Сведения об авторе Штумпф Артём Святославович — студент 3 курса бакалавриата, лаборант;
e-mail: artem.shtumpf@metalab.ifmo.ru.