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

***Понкратова Е.Ю., Штумпф А.С.***

*аспирант, студент*

*Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Физический факультет, Санкт-Петербург, Россия*

*ekaterina.grachkova@metalab.ifmo.ru, artem.shtumpf@metalab.ifmo.ru*

Обнаружение различных биологических соединений является трудоемким процессом из-за сложности их межмолекулярных связей [3]. Современные методы иммуноанализа и хроматографии не всегда позволяют добиться результатов в короткие сроки и с использованием небольшого количества ресурсов [1]. 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.

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.

Для изучения возможности анализа белковых соединений на первом этапе работы были получены спектры комбинационного рассеяния двух аминокислот (аланина и глутамина), а также их смесей в разных соотношениях. Было произведено сравнение рамановского сигнала смеси двух аминокислот с ожидаемыми спектрами, полученными путем сложения сигналов отдельных аминокислот с заданными коэффициентами. Спектры комбинационного рассеяния смесей аминокислот были обработаны с помощью алгоритма, основанного на нахождении минимума функции ошибок теоретически построенного спектра методом Лагранжа.

Далее методы машинного обучения были применены к различным задачам классификации и оценки возможности идентификации конкретных соединений по их рамановским спектрам:

* Двухклассовой классификации для сравнения спектров чистого аланина и глутамина, спектров смеси равных концентраций со спектрами дипептида.
* Трехклассовой классификации для сравнения сигналов от аланина, глутамина и их смеси в различных концентрациях.

Для этих задач были применены алгоритмы KNN, Random Forest [2]. Эти методы были применены к полученным данным в сочетании с перекрестной проверкой (cross validation), используемой для поиска лучших параметров моделей. Этот подход позволил получить как точность (precision), так и полноту (recall) выше 0,96 для набора данных, содержащего 2000 спектров.

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 3 и более аминокислот, что позволит анализировать сложные белковые соединения - гормоны.

**Литература**:

1. B. B. Hirpessa, B. H. Ulusoy, and C. Hecer, J. Food Qual. 2020, 10.1155/2020/5065386 (2020)
2. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, Inc., September 2019, 9781492032649.
3. Hunter, R., Anis, H. (2018). Genetic support vector machines as powerful tools for the analysis of biomedical Raman spectra. Journal of Raman Spectroscopy. doi:10.1002/jrs.5410.