

Анализ протеиногенных аминокислот и их пептидных соединений по спектрам комбинационного рассеяния с использованием машинного обучения

Штумпф А.С. (ИТМО)

Научный руководитель – Понкратова Е.Ю. (ИТМО)

Введение. Протеиногенные аминокислоты являются неотъемлемыми строительными элементами различных пептидных и белковых соединений, а возможные нарушения баланса аминокислот и соединений на их основе могут негативным образом повлиять на состояние человеческого организма. Современные методы обнаружения аминокислот и соединений на их основе, такие как высокоэффективная жидкостная хроматография[1], масс-спектрометрия[2] имеют свои недостатки, связанные со сложностями процедуры подготовки образцов и всей системы, вносящими существенные вариации в итоговые результаты[3-4]. Таким образом, научной проблемой, на решение которой направлена данная работа, является необходимость разработки быстрого и простого в использовании метода, позволяющего детектировать аминокислоты и соединения на их основе.

В данной работе было решено использовать спектроскопию комбинационного рассеяния - неинвазивный метод оптического биосенсинга - для детектирования аминокислот с последующим анализом полученных данных на основе машинного обучения[5]. Используемый в работе подход включает количественный и качественный анализ спектров комбинационного рассеяния аминокислот, пептидных соединений, а также применение алгоритмов машинного обучения для определения значений концентраций и идентификации компонент смесей аминокислот.

Основная часть. В работе были рассмотрены 20 протеиногенных аминокислот, их смеси в различных соотношениях, для которых были измерены спектры комбинационного рассеяния, ставшие основным объектом анализа, и составлена использованная в дальнейшем база данных. Решались две конкретные задачи: идентификация компонент смесей, соответствующая задаче классификации в машинном обучении, а также количественная характеристика образцов, то есть определение концентраций отдельных компонент, что соответствует задаче регрессии. Используемый в работе подход включал в себя многошаговую обработку исходных данных, применение и сравнение с помощью ряда метрик результатов для нескольких методов машинного обучения (KNN, Decision Tree, Random Forest, CNN, CatBoost) с использованием различного представления спектральных данных, отбор признаков и подбор гиперпараметров. Помимо этого, для повышения эффективности использования составленной базы данных была изучена возможность аугментации спектральных данных, заключающейся в использовании линейной комбинации спектров одиночных аминокислот, а также внесении шумов в имеющиеся данные.

Согласно полученным результатам, модели машинного обучения позволяют добиться точности 95% в рамках решения задачи идентификации аминокислот и 1% ошибки в задаче регрессии с использованием исключительно полученных экспериментальных данных. Помимо этого, были получены результаты решения поставленных задач для бинарных смесей с использованием аугментированных данных, демонстрирующие существенно отличающиеся зависимости для различных аминокислот. Также было рассмотрено влияние условий измерения спектров и сбора данных на стабильность сигнала и воспроизводимость результатов, что критически важно для обобщения исследуемого подхода на все протеиногенные аминокислоты.

Выводы. Результаты применения методов машинного обучения к полученным спектрам комбинационного рассеяния свидетельствуют о возможности получения точных

результатов при решении задач классификации и регрессии для смесей протеиногенных аминокислот. Используемая техника аугментации данных и подобранные оптимальные условия эксперимента позволяют продемонстрировать возможность распространения использованного подхода накратно большее число смесей аминокислот, то есть перейти к анализу более сложных систем, составными частями которых являются 3 и более аминокислот.

Список использованных источников:

1. Gałęzowska, G., Ratajczyk, J., & Wolska, L. (2021). Determination of amino acids in human biological fluids by high-performance liquid chromatography: critical review. *Amino Acids*, 53(7), 993. doi: 10.1007/s00726-021-03002-x
2. Mak, J., Cowan, T. M., & Le, A. (2019). Quantitative Analysis of Underivatized Amino Acids by Liquid Chromatography-Tandem Mass Spectrometry. *Methods Mol. Biol.*, 31347112. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/31347112>
3. Gao, L., Xu, P., & Ren, J. (2023). A sensitive and economical method for simultaneous determination of D/L- amino acids profile in foods by HPLC-UV: Application in fermented and unfermented foods discrimination. *Food Chem.*, 36628920. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/36628920>
4. Ferré, S., González-Ruiz, V., Guillaume, D., & Rudaz, S. (2019). Analytical strategies for the determination of amino acids: Past, present and future trends. *J. Chromatogr. B*, 1132, 121819. doi: 10.1016/j.jchromb.2019.121819
5. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, Inc., September 2019, 9781492032649.