ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ АНАЛИЗА ИК СПЕКТРОВ СЫВОРОТКИ КРОВИ ПАЦИЕНТОВ С ОНКОГЕМАТОЛОГИЧЕСКИМИ ЗАБОЛЕВАНИЯМИ

Бутяев Р.В.[1], Чернышев Д.А.[1], Михайлец Э.С.[1], Плотникова Л.В.[1], Гарифуллин А.Д.[2], Кувшинов А.Ю.[2], Волошин С.В.[2], Поляничко А.М.[1]

[1] Санкт-Петербургский государственный университет

ул. Ульяновская, 1, г. Петергоф, г. Санкт-Петербург, 198504, e-mail: st069812@student.spbu.ru

[2] Российский НИИ гематологии и трансфузиологии ФМБА России

ул. 2-я Советская, 16, г. Санкт-Петербург, 191024, РФ

Поступила в редакцию 24.07.2022. DOI: 10.29039/rusjbpc.2022.0545

Аннотация. Множественная миелома и хронический лимфолейкоз являются онкологическими заболеваниями крови, которые на сегодняшний день остаются неизлечимыми. В работе предложен метод классификации образцов сыворотки крови больных множественной миеломой, хроническим лимфолейкозом и здоровых доноров на основе анализа их спектров в среднем инфракрасном (ИК) диапазоне. ИК спектры сыворотки крови регистрировали с помощью ИК-Фурье спектрометра Tensor 27 в растворе D_2O. Для анализа полученных спектров в данной работе был реализован алгоритм машинного обучения – метод главных компонент. Использование метода главных компонент позволило существенно упростить представление массива спектральных данных. В работе проанализировали 45 образцов сыворотки крови. В результате применения данного подхода исследованный набор образцов разбивается на три непересекающихся множества, соответствующих образцам сыворотки крови больных множественной миеломой, хроническим лимфолейкозом и здоровых доноров. Таким образом, метод главных компонент может быть успешно применен для классификации образцов сыворотки крови пациентов с диагнозами множественная миелома и хронический лимфолейкоз. Универсальность предложенного алгоритма позволяет ожидать, что в будущем возможно применение аналогичного подхода и для других онкогематологических заболеваний.

Ключевые слова: *метод главных компонент, онкогематологические заболевания, ИК-спектроскопия, множественная миелома, хронический лимфолейкоз.*

Множественная миелома (ММ) и хронический лимфолейкоз (ХЛЛ) являются широко распространёнными и в настоящее время неизлечимыми онкологическим заболеваниями крови (4-5 случаев на 100000) [1]. ММ характеризуются накоплением злокачественных плазматических клеток, которые продуцируют избыточное количество моноклональных иммуноглобулинов [2]. Это, помимо прочего, приводит к тому, что в сыворотке крови изменяется соотношение двух самых распространённых белков - альбуминов и иммуноглобулинов. Так как иммуноглобулины и альбумины имеют разную вторичную структуру, то следить за изменением их соотношения возможно, анализируя ИК-спектры сыворотки крови, как это было продемонстрировано ранее [3,4].

Упомянутый подход, однако, требует индивидуальной, по сути, ручной, обработки огромного количества спектров, а потому плохо применим для скрининга ММ. В качестве альтернативы мы применили некоторые из методов машинного обучения, в частности, метод главных компонент (МГК) [5]. МГК позволяет упростить представление массива данных и изобразить этот массив в существенно меньшем по размерности пространстве главных компонент. Этот метод, в отличие от большинства методов машинного обучения, решает проблему “проклятия размерности” [6], то есть проблему экспоненциального роста сложности задачи по мере увеличения количества переменных. Это очень важно при обработке спектральных данных, размерность которых очень высока.

МАТЕРИАЛЫ И МЕТОДЫ

В работе использованы образцы сыворотки крови пациентов с диагнозами ММ и ХЛЛ, находящихся под наблюдением специалистов НИИ гематологии и трансфузиологии ФМБА России. Все образцы сыворотки крови были получены по описанной ранее методике [4].

В работе исследовали 45 образцов сыворотки крови: 12 образцов больных множественной миеломой, 10 образцов больных хроническим лимфолейкозом (ХЛЛ) и 23 образца здоровых доноров. Регистрацию спектров ИК поглощения проводили в растворах D_2O с использованием ИК Фурье-спектрометра Tensor 27 (Bruker). Спектральное разрешение составило 2 $см^{-1}$, представленные спектры получены путем усреднения по по 128-ми последовательным накоплениям. В работе использовали жидкостную разборную кювету с тефлоновым спейсером толщиной 50 мкм и окнами BaF_2. Для достижения полного изотопного замещения образцы сыворотки крови подвергали трехкратной лиофильной сушке с последующим перерастворением в D_2O (99.9% Sigma) согласно описанной ранее методике [7].

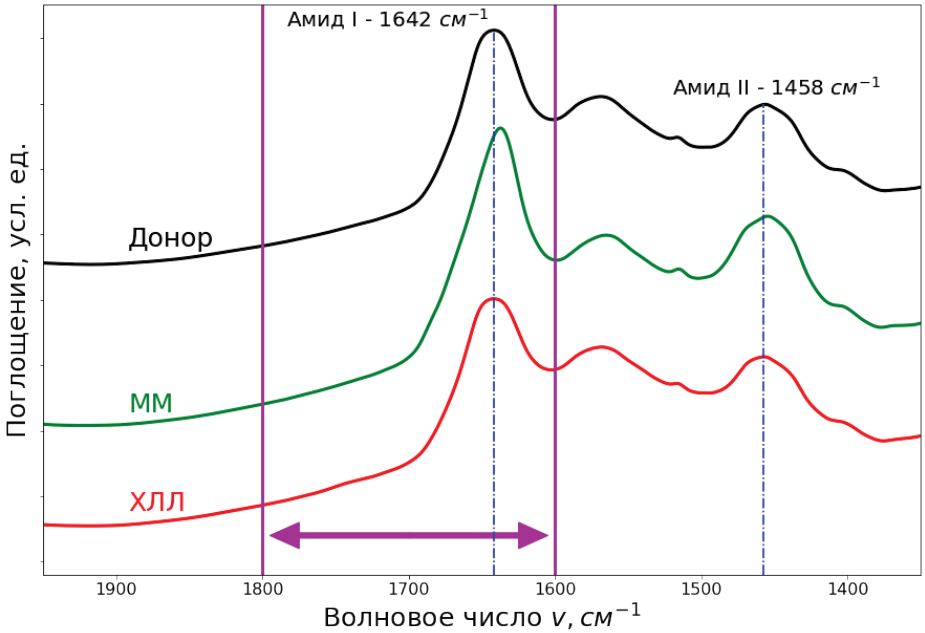

Рисунок 1. Характерные спектры образцов сыворотки крови людей трёх групп: ММ – множественная миелома, ХЛЛ – хронический лимфолейкоз, доноры. 1) - полосы поглощения амид I (пик – 1642 $см^{-1}$), отвечающие колебанию C=O связи. 2) – полосы поглощения амид II (пик – 1458 $см^{-1}$), отвечающие суперпозиции колебаний N-D и C-N связей. Стрелками выделен диапазон, к которому применялся МГК

МГК реализовали на языке Python (3.10) [8] в вычислительной среде Jupyter Notebook (версия 6.4.10-4) [9]. Программа написана на основе алгоритма NIPALS (нелинейный итеративный частичный метод наименьших квадратов) [5]. Перед применением алгоритма, спектры были нормированы и центрированы.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В ходе работы были получены и проанализированы ИК спектры сыворотки крови больных ММ, ХЛЛ и здоровых доноров. На рисунке 1 представлены характерные спектры образцов сыворотки крови больных ММ, больных ХЛЛ и здоровых доноров в диапазоне (1950-1350) $см^{-1}$. На всех спектрах наблюдаются полосы поглощения пептидной группы: полоса амид I (1600-1700) $см^{-1}$, отвечающая колебанию C=O связи и полоса амид II (1500-1400) $см^{-1}$, отвечающая суперпозиции колебаний N-D и C-N связей.

Анализируя ИК спектры белков, можно получить информацию об их вторичной структуре. Наиболее часто, для определения вторичной структуры белка, используется полоса амид I [7], однако другие полосы тоже могут нести важную информацию. Поэтому для анализа спектров мы применили метод главных компонент.

Для анализа полученных спектров, весь набор данных можно представить в виде матрицы X размера

$$M \times N,$$

где M – количество образцов; N – количество точек в спектре. МГК представляет спектральные данные (матрицу X) как набор линейно независимых векторов в новом пространстве. Эти вектора называются главными компонентами.

Согласно МГК [10] матрицу X можно разложить на две матрицы существенно меньшей размерности: матрицу T, которая является матрицей проекций точек на главные компоненты и называется матрицей счётов, и матрицу P, которая является матрицей главных компонент и называется матрицей нагрузок. При чём матрица счётов T, обладая намного меньшей размерностью относительно матрицы X, будет иметь практически всю информацию о спектральных данных.

$$X = TP^t \quad (1)$$

где T – матрица проекций точек на главные компоненты; P^t – транспонированная матрица, состоящая из главных компонент в старых координатах

Каждая из главных компонент отвечает за какой-либо скрытый параметр, характеризующий данные, при чём, чем старше главная компонента, тем за больший процент информации об исходной системе она отвечает.

Для оптимизации работы алгоритма необходимо сузить спектральный диапазон до области, содержащей информацию о вторичной структуре белка. Оптимального разделения точек удалось достичь при анализе диапазона 1800-1600 $см^{-1}$ (см. рис. 2).

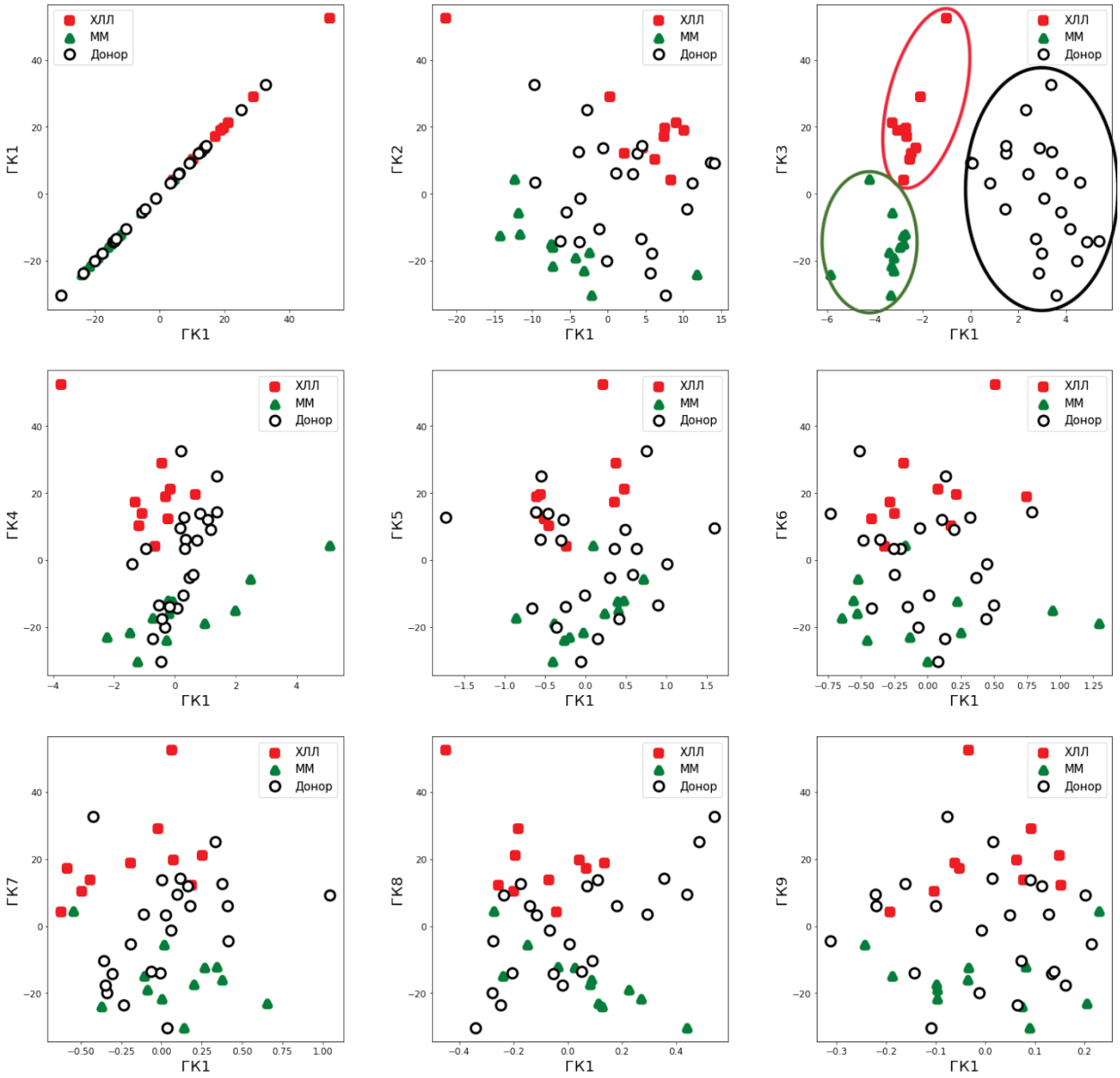

Рисунок 2. Зависимости первой главной компоненты от девяти первых главных компонент. Образцы на диаграмме обозначены следующим образом: ХЛЛ – красные квадраты, ММ – зеленые треугольники, здоровые доноры – черные окружности. На диаграмме ГК3-ГК1 точки, принадлежащие разным группам, выделяются наилучшим образом, поэтому группировки точек дополнительно обведены эллипсами соответствующего цвета – для наглядности

Применение МГК к выбранному диапазону дало следующие результаты. На рисунке 2 представлены первые девять главных компонент, отвечающие за 99,994% информации о системе.

Каждая из диаграмм представляет собой зависимость первой главной компоненты от одной из девяти главных компонент, а каждая точка отражает расположение одного образца в пространстве главных компонент. Для удобства, все точки раскрашены в один из трёх цветов, в зависимости от принадлежности образца: красные квадраты – больные ХЛЛ, зелёные треугольники – больные ММ, чёрные окружности– здоровые доноры.

Главные компоненты отражают поведение точек, вызванное каким-либо скрытым фактором, характеризующим данные. На всех диаграммах мы можем наблюдать по-разному распределённые точки. На диаграммах ГК4-ГК1 и ГК2-ГК1 наблюдается плотно лежащий набор точек, соответствующих образцам больных ХЛЛ. На диаграмме ГК4-ГК1 довольно плотно лежат точки ММ.

Однако наилучшим образом точки разделяются на диаграмме ГК3-ГК1. Там можно наблюдать три группы точек, соответствующих образцам ММ, ХЛЛ и здоровых доноров. Для удобства восприятия все точки выделены эллипсом цвета соответствующей группы.

Таким образом, опираясь на результат, представленный на диаграмме ГК3-ГК1, можно заключить, что МГК позволил разделить имеющиеся образцы на три группы в соответствии с диагнозом. Теперь, когда образцы из имеющегося набора классифицированы по трем группам, в соответствии с диагнозом, можно попробовать решить обратную задачу. Применим подход к спектру образца, не входившего в первоначальный, обучающий

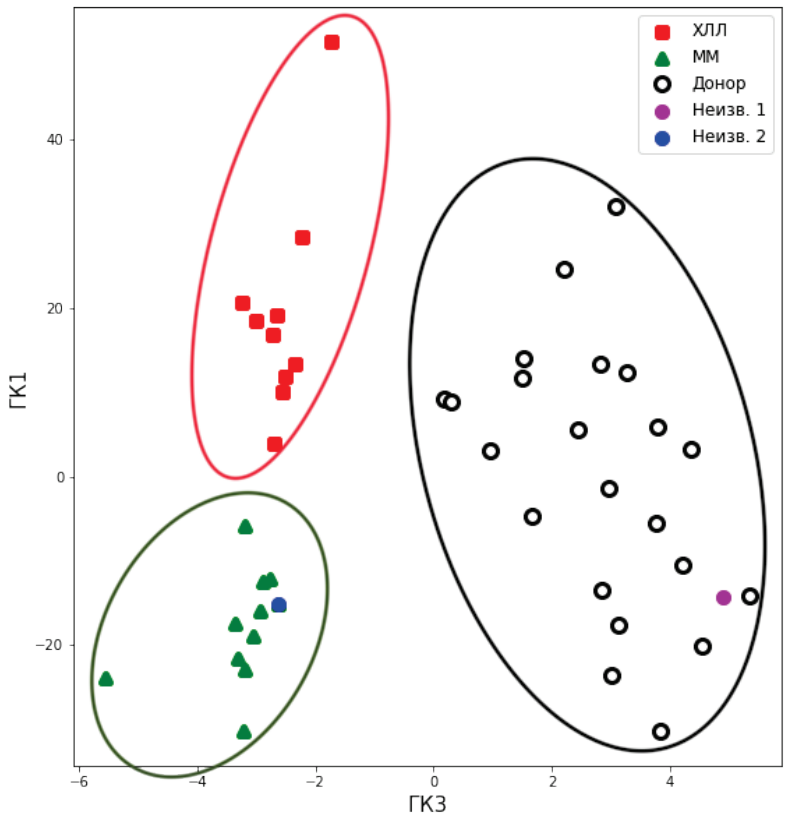

Рисунок 3. Зависимость первой главной компоненты от третьей. Образцы на диаграмме обозначены следующим образом: ХЛЛ – красные квадраты, ММ – зеленые треугольники, здоровые доноры – черные окружности, два случайных образца из набора: неизв. 1 - пурпурная точка, неизв. 2 – синяя точка

набор и рассмотрим его принадлежность к одной из групп. Для этого ИК-спектр такого образца X_{new} спроецируем в пространство главных компонент:

$$T_{new} = X_{new}\, P. \tag{2}$$

Мы получим матрицу счетов T_{new}. Отложив получившиеся счета в пространстве главных компонент на рисунке 2, мы сможем оценить принадлежность образца к той или иной группе по близости соответствующей точки к группам на диаграмме ГК3-ГК1.

Для примера, попробуем применить описанный алгоритм к двум случайным образцам. Результат показан на рисунке 3.

На рисунке 3 мы можем наблюдать, что один из случайных образцов (неизв. 1) попал в группу доноров, а второй (неизв. 2) попал в группу ММ. Отсюда, для обоих образцов можно дать оценку: образец неизв. 2, вероятно, взят у пациента, больного ММ, а образец неизв. 1 от здорового донора. Эта оценка совпадает с информацией о происхождении двух случайных образцов.

Таким образом, сочетание ИК-спектроскопии и МГК позволяет классифицировать спектры образцов сыворотки крови больных ХЛЛ и ММ и разделять их по группам. Такой подход в будущем, вероятно, можно будет применять и для других онкогематологических патологий, с целью их профилактики или раннего обнаружения признаков наличия заболеваний.

Часть работ выполнена с использованием оборудования Научного парка СПбГУ («Оптические и лазерные методы исследования вещества» «Центр диагностики функциональных материалов для медицины фармакологии и наноэлектроники», «Криогенный отдел»).

Список литературы / References:

1. Barlogie B., Gale R.P. Multiple Myeloma and Chronic Lymphocytic Leukemia: Commonalities and Differences in Biology and Therapy. *Leukemia & Lymphoma*, 1991, vol. 5, no. 1, pp. 27-32.

2. Raab M.S., Podar K., Breitkreutz I., Richardson P.G., Anderson K.C. *Multiple myeloma*. 2009, vol. 374, 16 p.

3. Mikhailets E.S., Chernyshev D.A., Telnaya E.A., Plotnikova L.V., Garifullin A.D., Kuvshinov A.Yu., Voloshin S.V., Polyanichko A.M. Protein secondary structure analysis of serum from patients with oncohematological diseases. *Journal of Physics: Conference Series*, 2021, vol. 2103, p. 012053, doi: 10.1088/1742-6596/2103/1/012053.

4. Тельная Е.А., Плотникова Л.В., Гарифуллин А.Д., Кувшинов А.Ю., Волошин С.В., Поляничко А.М. Инфракрасная спектроскопия сыворотки крови больных онкогематологическими заболеваниями. Санкт-

Петербургский государственный университет. *Биофизика*, 2020, т. 65, № 6, с. 1154-1160. [Telnaya E.A., Plotnikova L.V., Garifullin A.D., Kuvshinov A.Yu., Voloshin S.V., Polyanichko A.M. Infrared spectroscopy of blood serum of patients with oncohematological diseases. St. Petersburg State University. *Biophysica*, 2020, vol. 65, no. 6, pp. 1154-1160. (In Russ.)]

5. Dunn K. *Process Improvement Using Data*, 2010.
6. Powell W.B. Approximate Dynamic Programming: Solving the curses of dimensionality. *John Wiley & Sons*, 2007, vol. 703.
7. Поляничко А.М., Романов Н.М., Старкова Т.Ю. и др. Анализ вторичной структуры линкерного гистона H1 по спектрам инфракрасного поглощения. *Цитология*, 2014, т. 56, № 4, с. 316-322. [Polyanichko A.M., Romanov N.M., Starkova T.Y. et al.Analysis of the secondary structure of the linker histone H1 by infrared absorption spectra. *Tsitologiya*, 2014, vol. 56, no. 4, pp. 316-322. (In Russ.)]
8. *Python* 3.10.0.
9. Jupyter Notebook.
10. Родионова О.Е., Померанцев А.Л. *Хемометрика в аналитической химии*. 2006. [Rodionova O.E., Pomerantsev A.L. *Chemometrics in analytical chemistry*. 2006. (In Russ.)]

APPLICATION OF THE PRINCIPAL COMPONENT ANALYSIS TO IR SPECTRA OF BLOOD SERUM OF PATIENTS WITH ONCOHEMATOLOGICAL DISEASES

Butyaev R.V.[1], Chernyshev D.A.[1], Mikhailets E.S.[1], Plotnikova L.V.[1], Garifullin A.D.[2], Kuvshinov A.Yu.[2], Voloshin S.V.[2], Polyanichko A.M.[1]

[1] St. Petersburg State University

Ulyanovskaya str., 1, Peterhof, St. Petersburg, 198504, Russia; e-mail: robertmag@mail.ru

[2] Russian Research Institute of Hematology and Transfusiology of the FMBA of Russia.

2nd Sovetskaya str., 16, St. Petersburg, 191024, Russia

Received 24.07.2022. DOI: 10.29039/rusjbpc.2022.0545

Abstract. Multiple myeloma and chronic lymphocytic leukemia are oncological diseases of the blood, which remain incurable today. The paper proposes a method for classifying blood serum samples from patients with multiple myeloma, chronic lymphocytic leukemia and healthy donors based on the analysis of their spectra in the mid-infrared (IR) range. IR spectra of blood serum were recorded using a Tensor 27 IR Fourier spectrometer in D2O solution. To analyze the obtained spectra in this work, a machine learning algorithm was implemented – the principal component analysis. The use of the principal component analysis made it possible to significantly simplify the representation of the array of spectral data. 45 samples of blood serum were analyzed in the work. As a result of applying this approach, the studied set of samples is divided into three disjoint sets corresponding to blood serum samples of patients with multiple myeloma, chronic lymphocytic leukemia and healthy donors. Thus, the principal component method can be successfully applied to classify blood serum samples of patients with diagnoses of multiple myeloma and chronic lymphocytic leukemia. The universality of the proposed algorithm allows us to expect that in the future it is possible to apply a similar approach for other oncohematological diseases.

Key words: *Principal Component Analysis, oncohematological diseases, IR spectroscopy, multiple myeloma, chronic lymphocytic leukemia.*