КОМПЛЕКСНЫЙ АНАЛИЗ ФЛУКТУАЦИЙ ИНТЕНСИВНОСТИ ФЛУОРЕСЦЕНЦИИ МОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ

Н. Н. Яцков [1*], В. В. Скакун [1], Л. Недервин-Шипперс [2], А. Кортхольт [2], В. В. Апанасович [3]

УДК 535.37:547.96

[1] Белорусский государственный университет, 220030, Минск, Беларусь; e-mail: yatskou@bsu.by
[2] Университет Гронингена, 9747AG, Гронинген, Нидерланды
[3] Институт ИТ и бизнес-администрирования, 220004, Минск, Беларусь

(Поступила 4 февраля 2020)

Предложен метод комплексного анализа флуктуаций интенсивности флуоресценции молекулярных соединений, позволяющий определять структурный состав олигомеров белков. Идея метода состоит в анализе гистограмм числа фотоотсчетов экспериментальных измерений с применением метода главных компонент для оценки наличия олигомерных соединений и иерархического кластерного анализа для определения классов данных, соответствующих различным молекулярным соединениям, с последующим выделением медоидов кластеров для нахождения параметров олигомерного состава белковых комплексов. Эффективность алгоритмов анализа, разработанных в рамках реализации предложенного метода, подтверждена на смоделированных и экспериментальных гистограммах числа фотоотсчетов измерений флуктуаций интенсивности флуоресценции мономерных и димерных форм белка GFP.

Ключевые слова: *флуктуации интенсивности флуоресценции, гистограмма числа фотоотсчетов, молекулярные соединения, олигомеры белков, интеллектуальный анализ данных, метод главных компонент, иерархический кластерный анализ, зеленый флуоресцирующий белок GFP.*

A method is proposed for the complex analysis of fluctuations in the fluorescence intensity of molecular compounds, which allows determining the structural composition of protein oligomers. The idea of the method is to analyze the photon counting histograms of experimental measurements using principal component analysis to assess the presence of oligomeric compounds, and to perform hierarchical cluster analysis, to determine the data classes corresponding to various molecular compounds, followed by selecting cluster medoids to determine the oligomeric composition of protein complexes. The efficiency of the analysis algorithms developed within the framework of the proposed method was confirmed on simulated and experimental photon counting histograms of the measured fluorescence intensity fluctuations of monomeric and dimeric forms of green-fluorescent protein (GFP).

Keywords: *fluorescence intensity fluctuation, photon counting histogram, molecular compounds, protein oligomers, data mining, principal component analysis, hierarchical cluster analysis, green-fluorescent protein (GFP).*

Введение. Флуоресцентная флуктуационная спектроскопия широко используются для исследования диффузии белков и их взаимодействий в живых клетках [1—3]. В ходе эксперимента регистрируется флуоресценция молекул, связанных или свободно перемещающихся в растворе или клетке,

---

COMPLEX ANALYSIS OF FLUORESCENCE INTENSITY FLUCTUATIONS OF MOLECULAR COMPOUNDS

M. M. Yatskou [1*], V. V. Skakun [1], L. Nederveen-Schippers [2], A. Kortholt [2], V. V. Apanasovich [3]
(*[1] Department of Systems Analysis and Computer Modelling, Belarusian State University, Minsk, 220030, Belarus; e-mail: yatskou@bsu.by; [2] University of Groningen, 9747AG Groningen, The Netherlands; [3] Institute of IT&Business Administration, Minsk, 220004, Belarus*)

в некотором малом объеме (до 10^{-18} м^3), сформированном предельно сфокусированным лучом лазера. Флуктуации интенсивности флуоресценции обусловлены прежде всего изменением количества и местоположения молекул в регистрируемом объеме, а также их взаимодействием и свойствами среды. Олигомерный состав белкового соединения может быть определен путем анализа амплитуды флуктуаций интенсивности флуоресценции во времени (методы анализа распределения интенсивности флуоресценции — PCH (photon counting histogram) [4] и FIDA (fluorescence intensity distribution analysis) [5]). В методах PCH и FIDA для определения концентрации белка, свободно излучающего либо меченного люминесцентным красителем, строится гистограмма числа фотоотсчетов (ГЧФ) на заданном временном интервале регистрации. Регистрируемая интенсивность флуоресценции образца прямо пропорциональна количеству флуоресцирующих молекул, формирующих исследуемый молекулярный комплекс, что позволяет оценить количество молекул внутри белкового комплекса и размер комплекса [6, 7].

Для анализа распределения числа фотоотсчетов обычно используются различные математические модели [4—7] и методы оптимизации, среди которых наиболее часто применяется метод наименьших квадратов с оптимизацией Левенберга—Марквардта [8], позволяющий получить информацию о диффузионных и структурных свойствах исследуемых белковых соединений в первом приближении. Однако классические итерационные алгоритмы анализа данных имеют ряд существенных ограничений. Они не позволяют точно определить количество и вид молекулярных олигомеров, осуществляют локальный, а не глобальный поиск параметров моделей, требуют существенных вычислительных затрат на проведение анализа данных. Альтернативным подходом к решению указанной задачи является применение алгоритмов интеллектуального анализа и больших многомерных данных, суть которых состоит в одновременном глобальном анализе всего набора данных как единого целого [9—12].

В настоящей работе предложен метод комплексного анализа флуктуаций интенсивности флуоресценции и построенных на их основе ГЧФ с использованием алгоритмов интеллектуального анализа с целью определения олигомерного состава молекулярных соединений.

Методология. В основе разрабатываемого метода лежит гипотеза о разделимости набора многомерных экспериментальных данных в некотором информационном пространстве на несколько популяций, представляющих различные молекулярные олигомерные соединения [10]. Рассматривается малый объем регистрации, в котором в серии коротких интервалов времени преобладают молекулярные соединения одного вида. Предполагается нормальный закон распределений измеряемых характеристик для молекулярных соединений одного вида в выделенном пространстве. Например, мономеры белка могут формировать облако или сферообразный гауссов кластер данных в многомерном пространстве, построенном на основе измеряемых признаков. Если же к мономерным формам белка добавляются белковые олигомеры, то облако вытягивается или разделяется на две части вдоль некоторой линии, соединяющей центры двух популяций. В предельном случае ожидаются два облака или кластера данных мономеров и олигомеров. Таким образом, если группы данных разделяются на кластеры в многомерном пространстве признаков, то это является подтверждением наличия нескольких форм белковых соединений. Задачи подобного рода решаются с использованием алгоритмов интеллектуального анализа данных, таких как снижения размерности данных и кластерного анализа [10, 13, 14]. Алгоритмы снижения размерности позволяют перейти в пространство низкой размерности без потери сущности информации [15,16]. Алгоритмы кластерного анализа позволяют определить кластеры данных, заданные в той или иной мере сходства, число которых может быть связано с агрегатами молекулярных соединений. Так, использование метода главных компонент (МГК) позволит осуществить такое вращение, в результате которого ось первой главной компоненты совпадает с диагональю облака данных в многомерном пространстве [17]. Поэтому относительная доля разброса, приходящаяся на первую главную компоненту, для двух типов молекулярных соединений (ожидается вытянутый эллипсоид или два сферических облака данных в многомерном пространстве признаков) должна существенно отличаться от аналогичной для раствора мономеров (одно сферическое облако). Следует отметить информативность диаграммы рассеяния первых двух главных компонент в смысле определения структуры данных в двумерном пространстве.

Идея метода комплексного анализа состоит в подсчете ГЧФ на основе зарегистрированных интенсивностей флуоресценции (не исключено использование других характеристик, например автокорреляционной функции или факториальных кумулянтов распределения числа фотоотсчетов [18]), применении МГК для оценки наличия олигомерных соединений и иерархического кластерного ана-

лиза для определения групп данных, соответствующих различным молекулярным соединениям, с последующим выделением медоидов кластеров, ГЧФ, имеющих наименьшие средние расстояния до остальных объектов соответствующих кластеров, для оценки параметров олигомерного состава белковых комплексов. Для выполнения комплексного анализа требуется наличие экспериментальных данных для эталонного (мономеров) и тестируемого (олигомерных форм) образцов. Блок-схема разработанного метода представлена на рис. 1. Рассмотрим основные этапы метода.

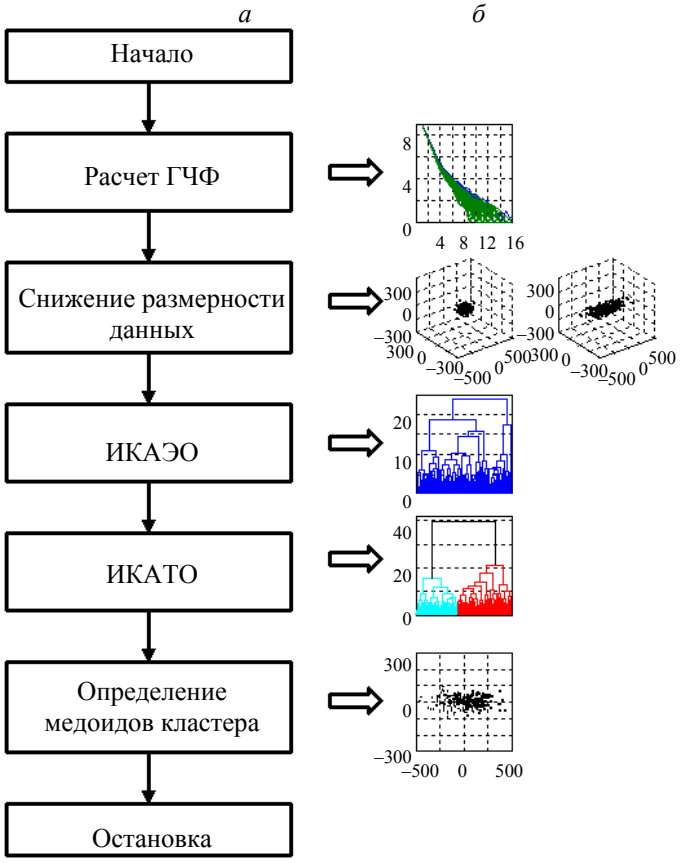

Рис. 1. Блок-схема метода (*а*) и диаграмма результатов его основных этапов (*б*) исследования флуктуаций интенсивности флуоресценции молекулярных соединений с использованием алгоритмов интеллектуального анализа данных

Расчет ГЧФ. Рассчитаем N ГЧФ на основе зарегистрированных наборов интенсивностей флуоресценции S_i, i = 1, 2, …, N, и сформируем объекты n_1, n_2, …, n_N, характеризуемые признаками $X_1, X_2, \ldots, X_K$, — каналами гистограмм, представляющими частоты появления f_j числа фотонов $l = (j - 1)$, j = 1, 2, …, K, в течение некоторого (короткого) интервала времени Δt. В качестве эталонного или референсного образца используем экспериментальные данные раствора мономеров, в качестве тестируемого образца — данные для олигомерных форм белка.

Снижение размерности данных. Применим МГК к наборам данных эталонного и тестируемого образцов. В МГК определяется такое линейное преобразование, в результате действия которого исходные данные X_1, X_2, …, X_K выражаются набором главных компонент Z_1, Z_2, ..., Z_K, где первые M главных компонент ($M<<K$), обеспечивают требуемую долю дисперсии γ групп признаков. В развернутом виде главная компонента Z_j выражается через векторы признаков X_1, X_2, …, X_K:

$$Z_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{Kj} X_K, \qquad (1)$$

где a_{ij} — параметры загрузки главных компонент. Относительная доля разброса (%), приходящаяся на главную компоненту Z_j:

$$\alpha_j = 100 \frac{D(Z_j)}{D(Z_1) + D(Z_2) + \ldots + D(Z_K)}, \qquad (2)$$

где $D(Z_j)$ — дисперсия компоненты Z_j. Если относительные доли разброса у эталонного и тестируемого образцов, приходящиеся на первую главную компоненту Z_1, не различаются, то считать, что олигомеры отсутствуют, означает остановить алгоритм. В ином случае — допустить наличие олигомеров и продолжить алгоритм.

Иерархический кластерный анализ эталонного образца (ИКАЭО). Выполним иерархический кластерный анализ гистограмм эталонного образца $n_1^Э$, $n_2^Э$, …, $n_N^Э$ в пространстве исходных признаков. В этом случае необходимо задать способ сравнения объектов между собой (или меру сходства, например евклидово, Минковского, корреляционное расстояния). В разработанном методе для устранения межэкспериментальных неоднородностей, связанных с раздельными измерениями эталонного и тестируемого образцов, предлагаем использовать стандартизированное евклидово расстояние (инвариантно к неоднородности в данных) [10]:

$$d_e(n_i, n_j) = \sqrt{\sum_{l=1}^{K} \frac{(x_{il} - x_{jl})^2}{\sigma_l^2}} \quad , \tag{3}$$

где x_{il} и x_{jl} — координаты объектов n_i и n_j; σ_l^2 — дисперсия признака X_l. Определим максимальное расстояние связи (или порог) d_1 на дендрограмме, при котором данные объединяются в один кластер. Максимальное расстояние связи d_1 используется в качестве порога для нахождения числа кластеров олигомеров на дендрограмме для тестируемых данных.

Иерархический кластерный анализ тестируемого образца (ИКАТО). Выполним иерархический кластерный анализ гистограмм тестируемого образца n_1^T, n_2^T, …, n_N^T в пространстве исходных признаков. По порогу d_1, найденному на предыдущем шаге алгоритма, выберем на дендрограмме кластеры данных. Полагаем, что один кластер принадлежит мономерам, другой(ие) — олигомерным формам.

Определение медоидов кластеров. На диаграмме рассеяния первых двух главных компонент отобразим кластеры мономеров и олигомеров. Сформируем наборы данных, вычислив медоиды в каждом кластере, для точного определения параметров молекулярных соединений с использованием методов PCH и FIDA.

Материалы и методы. Рассмотрим смоделированные и экспериментальные данные. Смоделированные позволяют качественно оценить работоспособность метода и исследовать пределы применения. Экспериментальные данные используются для подтверждения принципиальной возможности применения разработанного подхода к решению реальных задач экспериментальных исследований.

Имитационная модель потока фотоотсчетов с заданным распределением числа фотоотсчетов представлена в [19]. Количество фотонов, излученных молекулой за время наблюдения T, аппроксимируется распределением Пуассона с интенсивностью

$$\lambda_f = <q> TB(r), \tag{4}$$

где $<q>$ — яркость, или среднее число фотонов, излучаемое одной молекулой в единицу времени; $B(r)$ — функция профиля засветки; $r(x, y, z)$ — радиус-вектор молекулы. В качестве функции профиля засветки $B(r)$ используется трехмерное распределение Гаусса. Число молекул, находящихся в растворе в некотором объеме, подчиняется распределению Пуассона с параметром

$$\lambda_m = <N_m> V_0, \tag{5}$$

где $<N_m>$ — среднее число молекул исследуемого образца в единице объема; V_0 — объем засветки. Для каждой молекулы генерируются координаты расположения в объеме V_0 (по равномерному закону распределения) и число излученных фотонов (по распределению Пуассона с интенсивностью λ_f). Если моделируется смесь молекул разных видов, то необходимо выполнить циклы генерации фотонов для каждого вида молекул. Цикл генерации повторяется итерационно до накопления числа фотонов, при котором сформирована ГЧФ с заданным отношением сигнал/шум. Для учета эффекта разброса данных или “размытости” кластеров ГЧФ, обусловленных влиянием различных искажений, таких как наличие неустранимых примесей, тушащих или стимулирующих флуоресценцию молекул, высокий фоновый шум, засветка и деградация красителей, используется моделирование параметров модели, имеющих нормальное распределение с заданным математическим ожиданием и среднеквадратическим отклонением σ. Варьирование σ позволяет контролировать разброс данных или размытость кластеров кривых ГЧФ в многомерном пространстве временных отсчетов.

Смоделированные данные — пример идеализированной системы двух видов молекул: мономера (М) и димера (Д) некоторого белка (например, GFP в растворе), отдельно сгенерированные ГЧФ ко-

торых характеризуются средним числом молекул в объеме регистрации и их средней яркостью $<N^{М}> = 2$, $<q^{М}> = 5 \cdot 10^4$ и $<N^{Д}> = 1$, $<q^{Д}> = 10^5$. Интервал наблюдения $T = 5 \cdot 10^{-5}$ с. Проводилось моделирование с $\sigma = 0.02$ и 0.2 от абсолютных значений параметров $<N^{М}>$, $<q^{М}>$, $<N^{Д}>$ и $<q^{Д}>$.

Экспериментальные данные — хорошо известные мономерные и димерные формы зеленого флуоресцирующего белка GFP S65T [20] — предоставлены лабораторией биохимии клетки университета Гронингена (Нидерланды). Эталонные образцы: белок GFP в буферном лизисном растворе (50 мМ Tris, 50 мМ NaCl, 5 мМ DTT, 5 мМ $MgCl_2$, 1 % PI mix, 1 % Triton X-100); отдельные измерения мономера (mGFP) и стабильного димера (diGFP, синтезирован с помощью лигандирования вектора pDM313 в pDM334 в местах сайтов связывания SpeI/XBaI) белка GFP в лизатах клеток диктиостелиума. Тестируемый образец: смесь равных пропорций малых концентраций ($<N_m> <1$) белков mGFP и diGFP в лизате клеток диктиостелиума. Измерения первого образца выполнены с использованием флуоресцентного конфокального инвертированного микроскопа Leika TCS, оснащенного погруженным в масло объективом (100×, 1.4NA) и системой регистрации и записи потока фотоотсчетов PicoHarp 300 (PicoQuant). Второй и третий образцы исследованы с помощью сканирующего инвертированного конфокального микроскопа LSM 710 (Carl Zeiss), оснащенного погруженным в воду объективом (100x, 1.2NA) и системой измерения Confocor3 (Carl Zeiss). Флуоресценция образцов возбуждалась на $\lambda = 488$ нм, регистрировалась в полосе $\lambda = 505$—610 нм.

Смоделированные данные позволяют исследовать применимость разработанного метода в случае различной разделимости кластеров данных (варьируется с помощью параметра σ), соответствующих белковым соединениям. Данные, представляющие белок GFP в буферном растворе и лизате клеток, являются экспериментально подтвержденными и позволяют проверить работоспособность метода на примерах реальных модельных данных. Смесь мономерных и димерных форм белка GFP — пример набора данных, определенно содержащих различные формы агрегации белка. В предположении о нахождении в объеме наблюдения преимущественно молекул одного вида ГЧФ экспериментальных образцов строились на временном интервале $5 \cdot 10^{-2}$ с или меньше в одном измерении флуктуаций интенсивности флуоресценции длительностью 120 с.

Алгоритмы реализованы в среде математического программирования Matlab с применением функций pdist, linkage, cluster, eig, интегрирующих алгоритмы иерархического кластерного анализа и МГК [21]. Использован иерархический метод кластерного анализа, исследованы наиболее распространенные способ вычисления расстояния (стандартизированное евклидово) и мера сходства кластеров (Уорда) [13]. В МГК применена процедура центрирования данных. Для оценки ошибки ε восстановления ГЧФ различных видов молекул рассмотрено отношение неправильно определенных ГЧФ к общему числу ГЧФ (в %).

Результаты и их обсуждение. Результаты анализа смоделированных наборов данных с использованием алгоритмов комплексного подхода представлены на рис. 2 и в табл. 1. Выполнен анализ смоделированных данных раздельно для мономеров и димеров (рис. 2, *а* и *б*). Относительная доля разброса α_1, приходящаяся на первую главную компоненту, 54.6 и 58.8 % для мономеров и димеров, а облака данных в пространстве главных компонент имеют сферообразную гауссову форму. Пороговое значение меры схожести, при котором молекулы образуют единый кластер $d_1 = 15$, является критерием для определения кластеров различных форм молекул. Длина связи объединения результирующих кластеров в единый <2, что свидетельствует о значительной схожести объединяемых кластеров.

Применение алгоритмов разработанного метода к анализу объединенного набора смоделированных данных позволяет точно определить образцы мономерных и димерных форм белков (ошибка $\varepsilon = 0$), что подтверждается высокой относительной долей разброса, приходящейся на первую главную компоненту, $\alpha_1 > 98$ % (для мономеров 54.6 %), четкой разделимостью данных на два кластера в пространстве главных компонент Z_1 и Z_2 (рис. 2, *в*), высоким значением длины связи объединения результирующих кластеров в единый (>50), что подтверждает существенность различия кластеров. Следует отметить, что метод успешно работает в условиях рассмотренного примера размытости и частичного перекрытия кластеров данных ($\sigma = 0.2$, $\varepsilon = 1.5$ %; рис. 2, *г*), что характерно для молекулярных систем типа смеси мономера и димера белка GFP в лизате клеток. Определены образцы мономерных и димерных форм белков: относительная доля разброса $\alpha_1 = 99$ %, данные формируют два кластера в пространстве главных компонент Z_1 и Z_2 (рис. 2, *г*), длина связи объединения результирующих кластеров в единый >30.

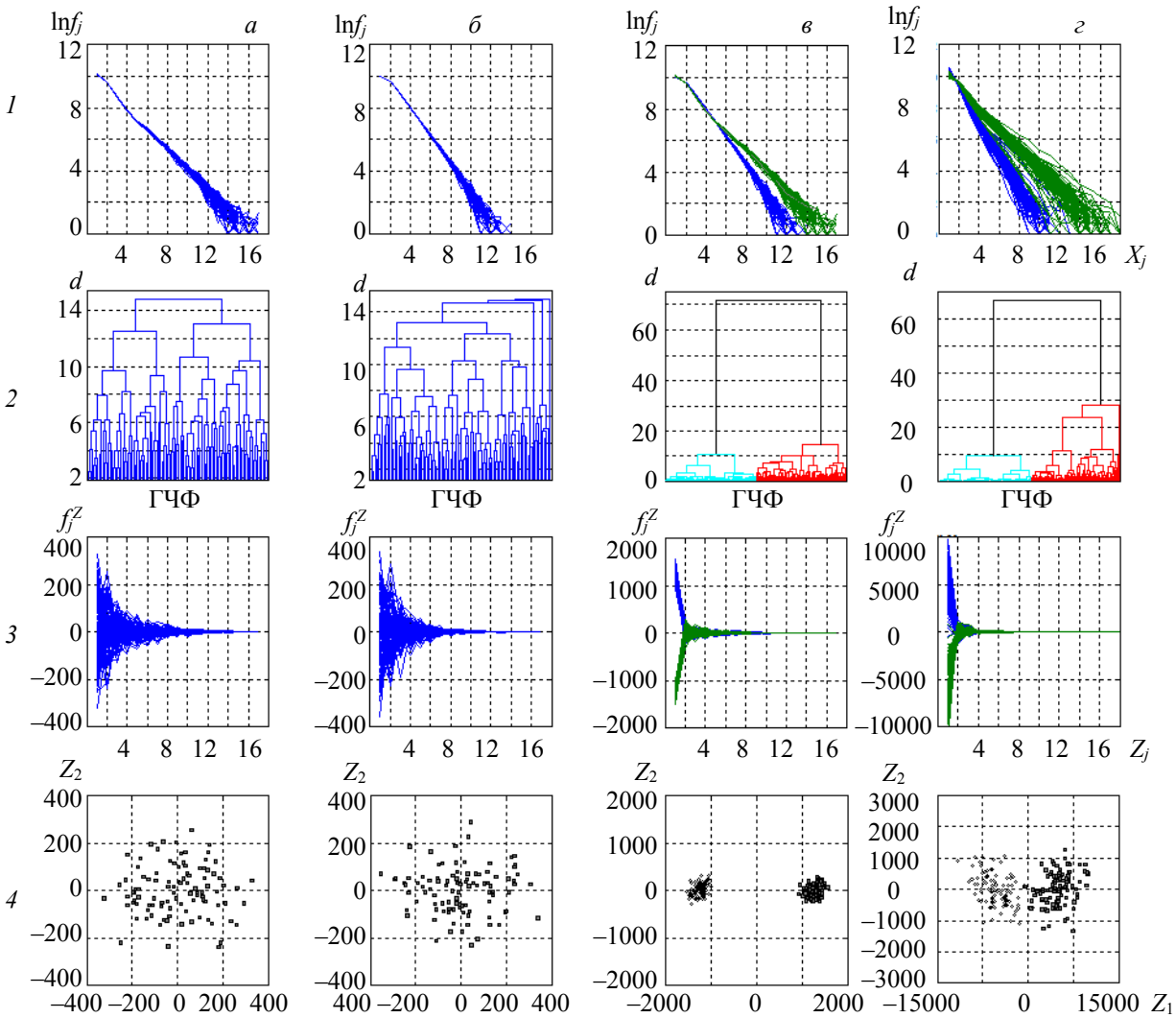

Рис. 2. Результаты анализа смоделированных данных с использованием разработанного метода, на основе алгоритмов метода главных компонент (выполнено центрирование данных) и иерархического кластерного анализа (реализованы стандартизированная евклидова мера сходства объектов и расстояние связи Уорда для объединения кластеров); параметры моделирования: $<N^{М}> = 2$, $<q^{М}> = 5 \cdot 10^4$ и $<N^{Д}> = 1$, $<q^{Д}> = 10^5$; *а* — мономеры, $\sigma = 0.02$; *б* — димеры, $\sigma = 0.02$; *в* и *г* — объединенные наборы мономеров и димеров с $\sigma = 0.02$ и 0.2; *1* — гистограммы счета фотонов в логарифмической шкале в пространстве исходных признаков $X_1, X_2, \ldots, X_K$; *2* — дендрограммы гистограмм счета фотонов ГЧФ, *d* — мера сходства кластеров; *3* — гистограммы счета фотонов в пространстве главных компонент $Z_1, Z_2, \ldots, Z_K$, f_j^Z — линейно преобразованные частоты появления числа фотонов в координатах главных компонент; *4* — гистограммы счета фотонов в пространстве двух первых главных компонент; размерность осей главных компонент — линейно преобразованные частоты появления числа фотонов в координатах компонент 1 и 2; оттенками серого цвета обозначены мономерные и димерные формы белков

В ходе исследования вместе со стандартизированным евклидовым расстоянием дополнительно рассмотрены три меры вычисления схожести между объектами, инвариантные к неоднородности данных, такие как Махаланобиса, корреляционное и Спирмена [9, 13, 14]. Наилучшие результаты получены для расстояний стандартизированного евклидова расстояния и Махаланобиса. Однако для меры Махаланобиса требуется вычисление ковариационной матрицы исходных данных, что может быть высокозатратным в случае анализа больших наборов данных ($N \to \infty$, $K \to \infty$).

Т а б л и ц а 1. Относительная доля разброса (в %), приходящаяся на первые 10 главных компонент, полученная в ходе анализа смоделированных (СД) и экспериментальных наборов данных с использованием метода главных компонент

Компоненты	1	2	3	4	5	6	7	8	9	10
СД, мономеры	**54.564**	29.263	8.134	3.079	2.318	1.120	0.701	0.471	0.166	0.094
СД, димеры	**58.775**	25.822	9.317	3.410	1.611	0.704	0.195	0.100	0.045	0.014
СД 1*	**98.768**	0.823	0.206	0.104	0.048	0.027	0.012	0.007	0.003	0.001
СД 2**	**98.998**	0.812	0.160	0.017	0.008	0.003	0.002	0.001	0.000	0.002
GFP	**50.502**	16.554	12.656	9.545	6.331	2.674	1.137	0.343	0.150	0.077
mGFP/diGFP	**99.869**	0.041	0.025	0.018	0.014	0.012	0.007	0.005	0.004	0.003
смесь mGFP/diGFP	**93.592**	4.161	1.360	0.470	0.175	0.104	0.055	0.028	0.023	0.011

* Мономеры/димеры, σ = 0.02;
** Мономеры/димеры, σ = 0.2.

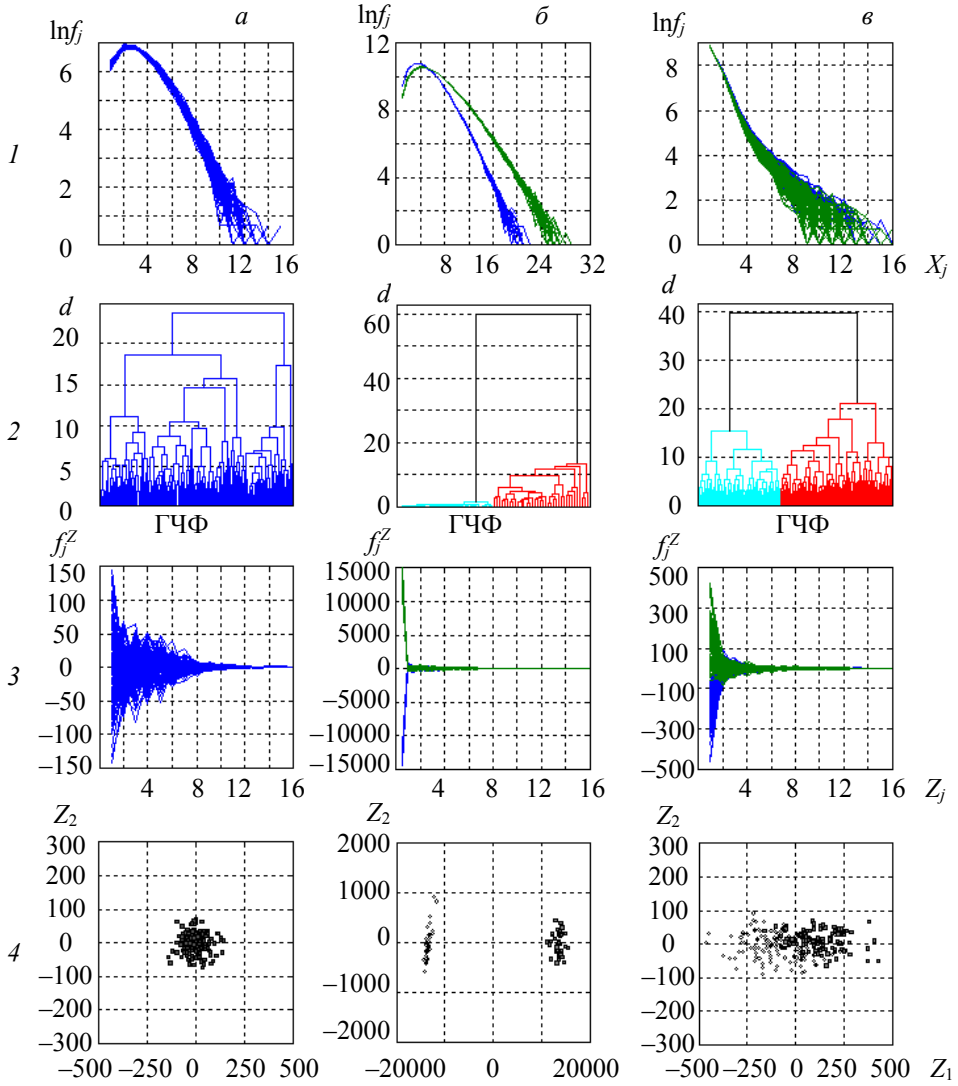

Рис. 3. Результаты анализа наборов экспериментальных данных с использованием разработанного метода, на основе алгоритмов метода главных компонент (выполнено центрирование данных) и иерархического кластерного анализа (реализованы евклидова мера сходства объектов и расстояние связи Уорда для объединения кластеров): *а* — белок GFP в буферном растворе, *б* — белки mGFP и diGFP в лизатах клеток, *в* — смесь белков mGFP и diGFP в лизате клеток; обозначения как на рис. 2

Результаты анализа экспериментальных наборов данных с использованием алгоритмов комплексного подхода представлены на рис. 3 и в табл. 1. Исследование данных для белка GFP в буферном растворе позволяет определить пороговое значение меры схожести ($d_1 = 23$), при котором мономеры образуют единый кластер, для использования в последующем анализе белковых соединений (рис. 3, *а*). Длина связи объединения результирующих кластеров в единый (<5), сферообразность облака данных в пространстве первых двух главных компонент (рис. 3, *а*) и невысокая относительная доля разброса $\alpha_1 = 50.5$ % (табл. 1), приходящаяся на первую главную компоненту, качественно подтверждают основополагающий принцип рабочей гипотезы, предложенной в реализованном методе. В результате анализа объединенных экспериментальных данных белков mGFP и diGFP в лизатах клеток подтверждено наличие двух форм белков, соответствующих мономерным и димерным формам (рис. 3, *б*): $\alpha_1 = 99.9$ %, данные формируют два кластера в пространстве главных компонент, длина связи объединения результирующих кластеров в единый >40. В ходе анализа экспериментальных данных смеси белков mGFP и diGFP в лизате клеток установлено наличие двух форм олигомеров белков. Относительная доля разброса α_1, приходящаяся на первую главную компоненту тестируемых данных, 93.6 % существенно превышает значение 50.5 %, полученное для мономерных форм белка GFP в буферном растворе. Расстояние связи, на котором формируется итоговый кластер, 40 (рис. 3, *в*), данные формируют два кластера в пространстве главных компонент, длина связи объединения результирующих кластеров в единый 18 существенно превосходит значение 5 для мономеров GFP. В качестве порогового значения для определения числа кластеров немономерных форм следует взять значение ≥23. При расстоянии связи 23 на дендрограмме тестируемых данных можно выделить два кластера, сформированных большинством молекул mGFP или diGFP (рис. 3, *в*). Дальнейшая оценка параметров белковых комплексов может быть проведена в ходе анализа медоидов полученных кластеров ГЧФ с использованием классических алгоритмов анализа данных флуоресцентной спектроскопии [5, 6]. Отметим, что мономеры белка GFP образуют сферообразный кластер данных в пространстве первых двух главных компонент (рис. 3, *а*), в то время как для смеси mGFP или diGFP наблюдается вытянутое эллипсоидальное облако, сформированное кластерами мономерной и димерной форм соединений (рис. 3, *в*).

Заключение. Предложен метод комплексного анализа флуктуаций интенсивности флуоресценции молекулярных соединений, позволяющий определить структурный состав олигомеров белков и дополняющий классические методы анализа PCH и FIDA. Эффективность алгоритмов, разработанных в рамках реализации предлагаемого метода, подтверждена в ходе анализа смоделированных и экспериментальных данных, представляющих флуоресценцию мономерных и димерных форм белка GFP. Разработанный метод имеет следующие преимущества над классическим методом анализа данных флуоресцентной флуктуационной спектроскопии: позволяет повысить точность анализа данных, так как использует весь набор данных, а не отдельные гистограммы; обеспечивает вычислительную производительность, обусловленную высокой скоростью выполнения процедур метода главных компонент и кластерного анализа в сравнении с отдельным анализом полного набора гистограмм; предоставляет возможность наглядной визуализации данных в пространстве первых двух главных компонент, что существенно информативнее диаграммы полного набора исходных гистограмм.

[1] **E. L. Elson, D. Magde.** Biopolymers, **13**, N 1 (1974) 1—27
[2] **A. Kitamura, M. Kinjo.** Int. J. Mol. Sci., **19**, N 4, pii: E964 (2018) 1—18
[3] **S. Veerapathiran, T. Wohland.** J. Biosci., **43**, N 3 (2018) 541—553
[4] **Y. Chen, J. D. Müller, P. T. So, E. Gratton.** Biophys. J., **77** (1999) 553—567
[5] **P. Kask, K. Palo, D. Ullmann, K. Gall.** Proc. Natl. Acad. Sci. USA, **96**, N 24 (1999) 13756—13761
[6] **Y. Chen, L. N. Wei, J. D. Müller.** Proc. Natl. Acad. Sci. USA, **100**, N 26 (2003) 15492—15497
[7] **В. В. Скакун, В. В. Апанасович.** Вестн. БГУ. Сер. 1, Физика. Математика. Информатика, **1** (2016) 52—59
[8] **D. Marquardt.** SIAM J. Appl. Math., **11**, N 2 (1963) 431—441
[9] **Н. Н. Яцков.** Интеллектуальный анализ данных, Минск, БГУ (2014)
[10] **Н. Н. Яцков, В. В. Скакун, В. В. Апанасович.** Прикладные проблемы оптики, информатики, радиофизики и физики конденсированного состояния, Минск, НИИ ПФП БГУ (2019) 122—124
[11] **M. Bramer.** Principles of Data Mining, Springer, London (2013)
[12] **C. C. Aggarwal.** Data Mining: The Textbook, Springer, eBook (2015)

[13] **И. Д. Мандель.** Кластерный анализ, Москва, Финансы и статистика (1988)
[14] **М. Б. Лагутин.** Наглядная математическая статистика, Москва, БИНОМ, Лаборатория знаний (2007)
[15] **P. V. Nazarov, A. K. Wienecke-Baldacchino, A. Zinovyev, U. Czerwińska, A. Muller, D. Nashan, G. Dittmar, F. Azuaje, S. Kreis.** BMC Med. Genom., **12**, N 1 (2019) 132(1—17)
[16] **N. Sompairac, P.V. Nazarov, U. Czerwinska, L. Cantini, A. Biton, A. Molkenov, Z. Zhumadilov, E. Barillot, F. Radvanyi, A. Gorban, U. Kairov, A. Zinovyev.** Int. J. Mol. Sci., **20**, N 18 (2019) E4414 (1—27)
[17] **I. T. Jolliffie.** Principal Component Analysis, Springer, New York (2002)
[18] **V. V. Skakun, E. G. Novikov, T. V. Apanasovich, V. V. Apanasovich.** Methods Appl. Fluores., **3**, N 4 (2015) 1—12
[19] **I. P. Shingaryov, V. V. Skakun, V. V. Apanasovich.** Methods Mol. Biol., **1076** (2014) 743—755
[20] **A. Kortholt, J. S. King, I. Keizer-Gunnink, A. J. Harwood, P. J. M. Van Haastert.** Mol. Biol. Cell, **18**, N 12 (2007) 4772—4779
[21] **Н. Н. Яцков, Е. В. Лисица.** Интеллектуальный анализ данных: методические указания к лабораторным работам, Минск, БГУ (2019)