

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ АНАЛИЗА ГИСТОГРАММ ЧИСЛА ФОТООТСЧЕТОВ В ЭКСПЕРИМЕНТАХ ФЛУОРЕСЦЕНТНОЙ ФЛУКТУАЦИОННОЙ СПЕКТРОСКОПИИ

В. В. Скакун ^{1*}, Н. Н. Яцков ¹, Л. Недервин-Шипперс ²,
А. Кортхолт ², В. В. Апанасович ¹

УДК 535.37

<https://doi.org/10.47612/0514-7506-2022-89-5-726-735>

¹ Белорусский государственный университет, Минск, Беларусь; e-mail: skakun@bsu.by

² Университет Гронингена, Гронинген, Нидерланды

(Поступила 16 июня 2022)

Для повышения эффективности анализа гистограмм счета фотонов при изучении молекулярного состава вещества методом флуоресцентной флукуационной спектроскопии предложен комплексный подход, основанный на применении методов интеллектуального анализа данных. Для проверки гипотезы о разделимости кластеров многомерных экспериментальных данных используется метод главных компонент. Исследована причина сжатия облака данных в характеристическую нелинейность, или так называемое дугообразное облако, в пространстве первых двух главных компонент. Рассмотрены примеры смоделированных наборов данных молекулярных систем различных яркостей и концентраций. Нелинейные эффекты усложняют интерпретацию и последующий количественный анализ данных. Установлено, что дугообразность облака данных — следствие наличия значительной вариации одного или нескольких физических параметров, в частности, результат значительного увеличения разброса яркости или концентрации молекул. Данные параметры могут служить дополнительными показателями при оценке качества экспериментов, если изучается только один тип молекулы, и при описании исследуемой системы — в случае смеси молекул разных видов. Предложено использовать нормировку локального взвешенного сглаживания диаграммы рассеяния для устранения нелинейных эффектов в пространстве главных компонент.

Ключевые слова: флуоресцентная флукуационная спектроскопия, гистограмма числа фотоотсчетов, имитационное моделирование, метод главных компонент, нормировка локального взвешенного сглаживания.

An integrated approach based on the use of data mining methods has been proposed to improve the efficiency of the analysis of photon counting histograms in the study of the molecular composition of a substance by the method of fluorescence fluctuation spectroscopy. The method of principal components is used to test the hypothesis about the cluster separability of multidimensional experimental data. The reason for the compression of a point cloud into a characteristic nonlinearity, or so-called arc-shaped cloud, in the space of first two principal components is investigated. Examples of simulated data sets on some selected molecular systems of various brightness and concentration are considered. Nonlinear effects complicate interpretation and subsequent quantitative analysis of data. It has been established that the arching of the data cloud is a consequence of the presence of a significant variation in one or more physical parameters. In particular, it is the result of a significant increase in the variation in the parameters of the brightness or concentration of molecules. These parameters can be as additional measure in assessing the quality of the experiments if only one type of molecule is studied, and also can be used for characterizing the system under study

PRINCIPAL COMPONENT ANALYSIS OF PHOTON COUNTING HISTOGRAMS IN FLUORESCENCE FLUCTUATION SPECTROSCOPY EXPERIMENTS

V. V. Skakun ^{1*}, M. M. Yatskou ¹, L. Nederveen-Schippers ², A. Kortholt ², V. V. Apanasovich ¹ (¹ Belarusian State University, Minsk, Belarus; e-mail: skakun@bsu.by; ² University of Groningen, Groningen, Netherlands)

in the case of a mixture of molecules of different types. It is proposed to apply the locally weighted scatterplot smoothing normalization to eliminate the nonlinear effects in the space of principal components.

Keywords: fluorescence fluctuation spectroscopy, photon counting histogram, simulation modelling, principal component analysis, local weighted scatterplot smoothing normalization.

Введение. Флуоресцентная флукутационная спектроскопия широко используется для исследования диффузии белков и их взаимодействий в живых клетках [1—3]. Перспективное направление исследований — разрешение смеси молекул разных видов на основе различия их характеристической яркости, определяемой числом излученных фотонов в единицу времени (cpms, counts per molecule per second), по гистограммам числа фотоотсчетов (ГЧФ) [2, 4]. Для анализа ГЧФ предложен ряд методов, среди которых наиболее популярны PCN (photon counting histogram) [4], FIDA (fluorescence intensity distribution analysis) [5] и их специальные модификации [6—9]. Для повышения эффективности анализа данных применяется совместный анализ автокорреляционных функций и ГЧФ — BDGA (brightness and diffusion global analysis) [10, 11].

Для оценки гетерогенности образца в наборе экспериментальных измерений предложен комплексный подход на основе применения методов интеллектуального анализа данных [12]. Ключевым элементом подхода является использование метода главных компонент (МГК) [13] для проверки гипотезы о разделимости набора многомерных экспериментальных данных на несколько поднаборов, соответствующих различным видам молекул, в пространстве первых главных компонент (ГК). МГК — наиболее простой и часто весьма эффективный метод снижения размерности данных, широко применяемый, например, при анализе ИК-Фурье-спектров [14], использует линейное преобразование и относительно прост в программной реализации. Диаграммы МГК, построенные по спектрам или различного рода характеристикам флуоресценции молекулярных систем (ГЧФ, автокорреляционным функциям потока фотоотсчетов, кривым затухания флуоресценции), позволяют определить схожесть или различие образцов и протекающих в них биофизических процессов. Предполагается, что ГК флукутаций интенсивности флуоресценции исследуемых молекулярных систем обусловлены влиянием двух независимых факторов — яркостью и концентрацией молекулярных соединений [4]. Для набора данных некоторой молекулярной системы в состоянии термодинамического равновесия ожидается сферообразный или эллипсоидальный кластер ГЧФ в пространстве первых ГК [12]. Однако для ряда исследуемых систем, например белков GFP и diGFP в цитоплазме “истощенных” клеток диктиостелиума, наблюдается дугообразное облако (рис. 1, а), не объяснимое факторами яркости и числа молекул, что затрудняет применение методов определения видов молекул в интегрированных данных.

Схожий вид диаграммы ГК получен в результате МГК-анализа данных белка FKBP12 (FK506 binding protein 12)-GFP [11] (рис. 1, б).

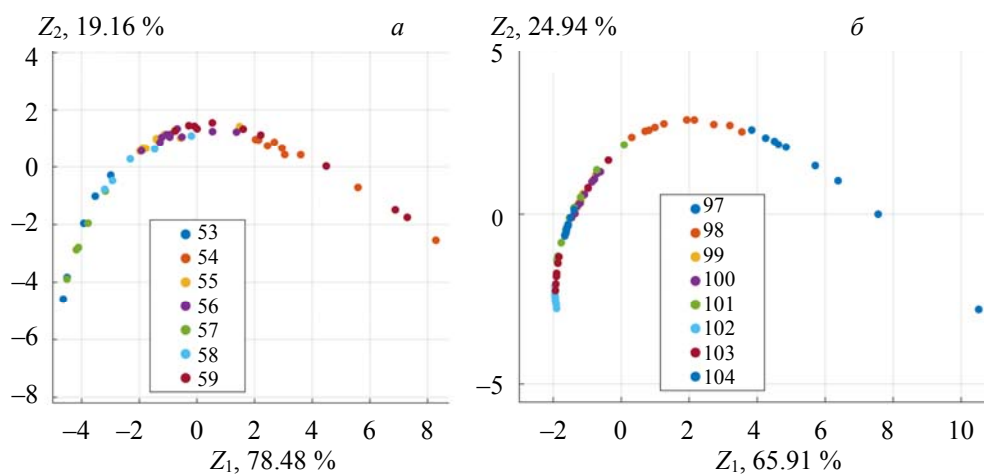


Рис. 1. Диаграммы МГК первых двух главных компонент, полученные в результате анализа наборов данных ГЧФ белков GFP и diGFP (а) и FKBP12-GFP (б) в цитоплазме “истощенных” клеток диктиостелиума, указан номер клетки в экспериментальном наборе данных

Параболическая форма облака данных предполагает наличие статистической взаимосвязи между молекулярной яркостью и концентрацией. Потенциально статистическая взаимосвязь числа молекул и их характеристической яркости может наблюдаться в системах, в которых происходит непрерывный процесс димеризации, а время регистрации ГЧФ намного меньше скорости димеризации: в подобном процессе две молекулы яркости x могут формировать димер яркостью $2x$. Рассматриваем наборы ГЧФ, полученные независимо для мономеров и димеров GFP. Процесс димеризации и распада димеров предполагается незначительным. Процесс фотообесцвечивания также может воздействовать на количество молекул и их яркость. Например, если молекула GFP, входящая в состав димера, перестает излучать фотоны, то димер первоначально регистрируется как мономер, затем перестает быть видимым. Однако неясно, вызывают ли указанные эффекты нелинейность ГК. Кроме того, сжатие облака приводит к деформации кластеров данных, что существенно усложняет их интерпретацию. Поэтому для эффективного применения МГК к рассмотренным примерам наборов данных необходимо разработать метод, способный устранить эффект нелинейности данных в пространстве первых ГК.

В настоящей работе исследованы причины появления нелинейного эффекта в пространстве первых двух ГК на примерах искусственно смоделированных наборов данных, где степень детализации изучаемых процессов предельно высока и априори известна. Устранение нелинейных эффектов выполнено с помощью нормализации локального взвешенного сглаживания данных (lowess – local weighted scatterplot smoothing) [15].

Материалы и методы. Хорошо известные мономерная и димерная формы зеленого флуоресцентного белка GFP S65T [16] предоставлены лабораторией биохимии клетки университета Гронингена (Нидерланды). Проведены отдельные измерения мономера (mGFP) и стабильного димера (diGFP, синтезированного лигандированием вектора pDM313 в pDM334 в сайтах связывания SpeI/XbaI) белка GFP в живых клетках диктиостелиума (*Dictyostelium*). Процедура подготовки образцов клеток описана в [11]. Образцы исследованы с помощью сканирующего инвертированного конфокального микроскопа LSM 710 (Carl Zeiss), оснащенного погруженным в воду объективом (100×, 1.2NA) системой измерения ConfoCor 3 (Carl Zeiss). Флуоресценция образцов возбуждалась на $\lambda = 488$ нм, регистрировалась в полосе $\lambda = 505$ —610 нм.

Метод главных компонент. Идея МГК состоит в учете корреляционной связи между признаками $X_1, X_2, \dots, X_K$, где K — координаты исходного пространства признаков, а именно каналы гистограмм, представленные частотами появления $f_j, j = 0, 1, 2, \dots, K$, того или иного числа фотонов в течение некоторого (короткого) интервала времени Δt , на основе которой можно совершить переход в пространство новых переменных Z [13]. Линейное вращение многомерной системы координат данных X осуществляется таким образом, чтобы дисперсия новых признаков перераспределилась в сторону уменьшения. Несмотря на то что суммы дисперсий в исходном X и новом Z пространствах признаков равны, вклад большей части компонент в объясняемую дисперсию оказывается небольшим. Исключают из рассмотрения компоненты с малым вкладом. С помощью первых (наиболее весомых) ГК Z_1 и Z_2 можно объяснить основную часть суммарной дисперсии в данных. Достоинством МГК является сохранение геометрических расстояний между объектами в пространстве ГК [17].

Нормировка локального взвешенного сглаживания данных. Идея метода состоит в нахождении некоторой сглаживающей функции или нелинейной регрессионной модели f_{lowess} , аппроксимирующей дугообразное облако экспериментальных данных, и последующем преобразовании ГК Z_2 с использованием восстановленной регрессионной модели, в результате которого набор ординат распределен вдоль некоторой прямой, параллельной оси абсцисс (дугообразное облако преобразуется в эллипсоидальное, главная ось которого коллинеарна оси абсцисс) [17]:

$$Z'_2 = Z_2 f_{\text{lowess}} \quad (1)$$

Имитационное моделирование систем флуоресцентной флуктуационной спектроскопии. Модель потока фотоотсчетов представлена в [18] и позволяет смоделировать наборы ГЧФ флуктуаций флуоресценции молекулярных соединений, различающихся средним числом молекул и их яркостью. Количество фотонов, излученных молекулой за время наблюдения T , аппроксимируется распределением Пуассона с параметром

$$\lambda_f = \langle q \rangle TB(\mathbf{r}), \quad (2)$$

где $\langle q \rangle$ — яркость, или среднее число фотонов, излучаемое одной молекулой в единицу времени, $B(\mathbf{r})$ — функция профиля засветки, $\mathbf{r}(x, y, z)$ — радиус-вектор молекулы. В качестве функции профи-

ля засветки $B(\mathbf{r})$ используется асимметричное трехмерное, вытянутое вдоль оси z распределение Гаусса:

$$B(\mathbf{r}) = B(x, y, z) = B_0 \exp(-2(x^2 + y^2)/\omega_0^2 - 2z^2/z_0^2), \quad (3)$$

где ω_0 и z_0 — размер профиля в латеральном и аксиальном направлениях.

Число молекул, находящихся в растворе в некотором открытом объеме, подчиняется распределению Пуассона с параметром

$$\lambda_m = \langle N \rangle V_0, \quad (4)$$

где $\langle N \rangle$ — среднее число молекул исследуемого образца в единице объема; V_0 — объем засветки. Внефокусное излучение моделируется по распределению Пуассона с параметром [18]:

$$\lambda_{\text{of}} = \langle N \rangle \langle q \rangle T F / 2\sqrt{2}(1 + F)^2, \quad (5)$$

где F — величина внефокусной засветки.

Алгоритм моделирования следующий. Для каждой молекулы генерируются координаты расположения в объеме V_0 (по равномерному закону распределения) и число излученных фотонов (по распределению Пуассона с параметром λ_j). 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 естественной вариацией физических параметров, а также влиянием различных искажений, таких как наличие неустраняемых примесей, тушащих или стимулирующих флуоресценцию молекул, высокий фоновый шум и деградация красителей, моделируется дополнительный разброс параметров модели по нормальному распределению с заданным математическим ожиданием и среднеквадратическим отклонением (СКО) σ . Варьирование σ позволяет контролировать разброс данных или размытость кластеров кривых ГЧФ в многомерном пространстве временных отсчетов.

Смоделированные данные — примеры идеализированных систем видов флуоресцирующих молекул: мономеров, димеров и агрегатов некоторого белка, помеченного флуоресцирующей меткой, или белка, обладающего собственной флуоресценцией, например GFP, отдельно сгенерированные ГЧФ которых характеризуются параметрами $\langle N \rangle$ и $\langle q \rangle$. Имитационное моделирование позволяет исследовать эффекты дугообразности облаков ГЧФ в пространстве ГК при изменении параметров $\langle N \rangle$, $\langle q \rangle$ и их разброса (варьируется с помощью σ), соответствующих различным экспериментальным системам, таким как белковые соединения различной яркости и концентрации.

Имитационное моделирование затухания флуоресценции молекул. Для выяснения причин появления дуги в пространстве ГК проведено имитационное моделирование совершенно иной, предельно упрощенной молекулярной системы. В качестве такой системы можно рассмотреть некоторый идеализированный флуоресцирующий образец, спонтанно испускающий фотоны под воздействием периодических ультракоротких вспышек лазера с длиной волны возбуждения в области максимума спектра поглощения. Регистрируя процесс флуоресценции по методу счета единичных коррелированных по времени фотонов (time-correlated single photon counting) и полагая инструментальную функцию отклика прибора регистрации флуоресценции дельта-импульсом $e(t)$, экспериментальную кривую затухания флуоресценции $E(t)$ аппроксимируем одноэкспоненциальной функцией $F(t, a, \tau)$ [19]:

$$F(t, a, \tau) = e(t) \otimes f(t, a, \tau) = \int_0^t e(t-x) f(x, a, \tau) dx, \quad (6)$$

где

$$f(t, a, \tau) = a \exp(-t/\tau), \quad (7)$$

a — предэкспоненциальный фактор; τ — время затухания флуоресценции.

Имитационная модель такой системы хорошо известна и представляет собой генерацию интервалов времени между вспышкой лазера и моментом регистрации фотона по экспоненциальному распределению [20]. Гистограмма полученных времен представляет собой кривую затухания флуоресценции. Генерации повторяются итерационно до накопления числа фотонов, при котором сформирована кривая затухания флуоресценции с заданным отношением сигнал/шум.

Описание вычислительного эксперимента. Для имитационного моделирования ГЧФ исследованы молекулярные системы 1 и 2 — растворы некоторого свободно диффундирующего флуоресцирующего белка, например GFP, в буфере при различных концентрациях и характеристической яркости. Процесс формирования димеров и агрегатов моделируется увеличением характеристической яркости

в целое число раз (для димеров в 2 раза, для оптамеров в 4 раза). Результаты имитационного моделирования без потери общности можно распространить на более сложные системы, такие как GFP в лизате или цитоплазме живых клеток.

Моделирование осуществлялось следующим образом. Если моделировалась простая однокомпонентная система, например только мономеры GFP, то генерировалось 100 реализаций ГЧФ для одного набора параметров с последующим МГК-анализом полученного набора данных. Моделирование набора ГЧФ, составленного из измерений разных молекулярных систем, например мономеров, димеров и оптамеров, включало в себя 50 независимых генераций реализаций ГЧФ для одного вида молекулярной системы с последующим МГК-анализом всего полученного набора данных. Прочие параметры модели: внефокусное излучение $F = 0.8$, интервал наблюдения $T = 5 \cdot 10^{-5}$ с. Проведено 10^5 циклов генераций для моделирования одной ГЧФ. Суммарный временной интервал регистрации флуктуаций флуоресценции $5 \cdot 10^{-5} \cdot 10^5 = 5$ с. Параметры моделирования выбирались такими, чтобы максимально исследовать данный эффект.

Система 1 представляет собой несколько видов идеализированных молекул, характеризующихся средним числом молекул $\langle N \rangle = 1, 2, 5$ в объеме регистрации и яркостью $\langle q \rangle = 5, 10$ и $20 \cdot 10^4$ crms. Это пример сравнительно простой системы, где среднее число молекул постоянно, а яркость молекул изменяется, как в случае измерения флуктуаций интенсивности флуоресценции мономеров, димеров и оптамеров некоторого флуоресцирующего белка или белка, помеченного флуоресцирующими метками, или яркость постоянна, а среднее число молекул (концентрация белка) изменяется. Для исследования причины появления нелинейности при трансформации данных дополнительно проводится моделирование системы с варьируемым параметром σ в распределении параметров яркости и среднего числа молекул. Данные эксперименты дополнительно учитывают эффекты, связанные со значительным увеличением вариации в измеряемых параметрах яркости и концентрации, что часто связано с различными экспериментальными искажениями (фотообесцвечиванием, нестабильностью лазера, деградацией молекул, движением внутриклеточных структур) и неоднородностью образца. Исследуется влияние яркости/концентрации молекул и разброса в параметрах молекулярных систем на диаграмму рассеяния ГК.

Система 2 включает в себя один вид идеализированных молекул, характеризующихся постоянными средним числом молекул $\langle N \rangle = 1$ и яркостью $\langle q \rangle = 1 \cdot 10^5$ crms. Для исследования причины появления нелинейных эффектов дополнительно проводится моделирование с вариацией σ . Это пример простой “эталонной” системы, где дополнительно учитываются эффекты, связанные со значительным увеличением вариации в измеряемых параметрах яркости, что часто связано с различными экспериментальными искажениями, такими как неоднородность образца или низкое качество проведения экспериментов флуоресцентной флуктуационной спектроскопии.

Система 3 представляет собой тот же белок GFP, медленно диффундирующий в существенно неоднородной среде (среда может оказывать значительное влияние на флуоресцентные свойства), или иммобилизованный белок на мембране (или другой поверхности), но также подверженный разнообразным процессам, влияющим на время затухания флуоресценции, рассмотрена для имитационного моделирования кривых затухания флуоресценции. Параметры системы: предэкспоненциальный фактор $a = 0.6$, время затухания флуоресценции $\tau = 2$ нс. Ширина канала и число каналов в гистограммах 0.012 нс и 1024. Моделировалось 100 кривых затухания, содержащих $5 \cdot 10^5$ фотоотсчетов. Время затухания флуоресценции дополнительно варьировалось по определенному закону распределения с установленным СКО.

Моделирование и вычислительные эксперименты проведены в среде MATLAB. Вычислительные процедуры алгоритмов уменьшения размерности данных выполнены с помощью MATLAB-функции `rca`. Lowess-нормализация реализована в среде статистического программирования R с применением функции `lowess` пакета `stats` [21]. В МГК использована процедура центрирования и шкалирования данных.

Результаты и их обсуждение. Как и ожидалось, для исследуемых систем 1—3 с фиксированными параметрами при малых σ в пространстве ГК отчетливо формировались сферические или эллипсоидальные кластеры точек. С увеличением разброса некоторого параметра моделируемой системы кластеры вытягивались и сжимались в характерную дугу. Система 1 является прямой моделью молекулярной системы, представленной на рис. 1, *a*. Моделирование системы без дополнительного разброса концентрации и яркости молекул демонстрирует хорошо разделяемые облака точек в пространстве первых двух ГК [12] (рис. 2, *a*).

Увеличение разброса (σ) по яркости и концентрации, характеризуемой средним числом молекул в единице объема, по распределению Гаусса со средним, равным моделируемому параметру и изменяющимся от незначительных до больших σ , приводит к вытягиванию облаков данных и их дугообразности. Результаты применения МГК к сгенерированным наборам данных приведены на рис. 2, *б*—*д*.

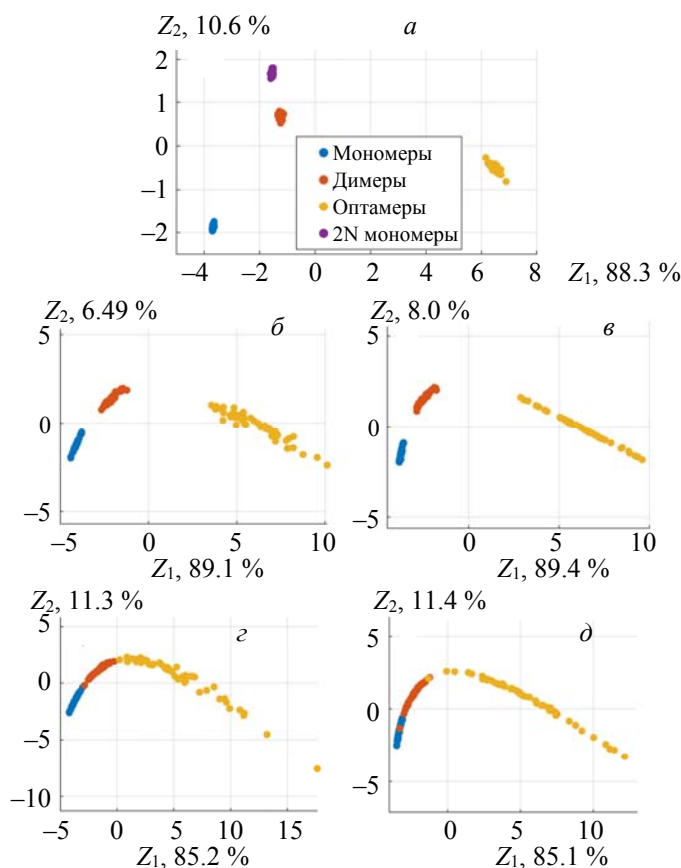


Рис. 2. Диаграммы МГК первых двух главных компонент, полученные в результате анализа наборов данных ГЧФ для системы 1: *a* — без разброса параметров системы; *б* — относительно небольшой разброс по яркости ($q_1 = 5 \cdot 10^4$, $\sigma_1 = 2 \cdot 10^3$; $q_2 = 1 \cdot 10^5$, $\sigma_2 = 5 \cdot 10^3$; $q_3 = 2 \cdot 10^5$, $\sigma_3 = 1 \cdot 10^4$) при постоянной концентрации ($N_1 = N_2 = N_3 = 1$, $\sigma = 0.05$); *в* — относительно небольшой разброс по концентрации ($N_1 = 1$, $\sigma = 0.075$; $N_2 = 2$, $\sigma = 0.15$; $N_3 = 5$, $\sigma = 0.3$) при постоянной яркости ($q_1 = q_2 = q_3 = 1 \cdot 10^5$, $\sigma_1 = 1 \cdot 10^3$); *г* — значительный разброс по яркости ($q_1 = 5 \cdot 10^4$, $\sigma_1 = 7.5 \cdot 10^3$; $q_2 = 1 \cdot 10^5$, $\sigma_2 = 1.5 \cdot 10^4$; $q_3 = 2 \cdot 10^5$, $\sigma_3 = 3 \cdot 10^4$) при постоянной концентрации ($N_1 = N_2 = N_3 = 1$, $\sigma = 0.05$); *д* — значительный разброс по концентрации ($N_1 = 1$, $\sigma = 0.15$; $N_2 = 2$, $\sigma = 0.3$; $N_3 = 5$, $\sigma = 0.75$) при постоянной яркости ($q_1 = q_2 = q_3 = 1 \cdot 10^5$, $\sigma_1 = 1 \cdot 10^3$)

В исследуемых системах в пространстве ГК наблюдается ожидаемое уширение кластеров точек, переходящее в дугу. Вид дуги типичен для случая вариации как яркости, так и концентрации. Результат моделирования изучаемых систем позволяет сделать предположение, что причина нелинейности — увеличение дисперсии физических параметров системы. Форма дуги определяется распределением определенных параметров и не зависит в первом приближении от вида физического эксперимента. Кроме того, очевиден обратный вывод: независимо от закона распределения параметра дугообразность проявляется при увеличении СКО некоторого параметра системы. Для подтверждения данной гипотезы исследована принципиально иная, упрощенная система 3. Выбор моделируемой системы проведен таким образом, чтобы исключить возможные статистические корреляции параметров модели, приводящие к формированию неэллипсоидальных кластеров в пространстве ГК. Сгенерированы кривые затухания флуоресценции с параметрами $a = 0.6$ и $\tau = 2$ нс, число каналов гистограммы 1024, ширина канала 0.012 нс. Исследованы три вида распределения основного параметра модели — вре-

мени затухания флуоресценции τ при изменении σ : нормальное со средним, равным τ , и заданным σ , равномерное в некоторой окрестности τ и показательное с параметром, равным некоторой доли от τ . Результаты применения МГК к набору сгенерированных кривых затухания флуоресценции показаны на рис. 3.

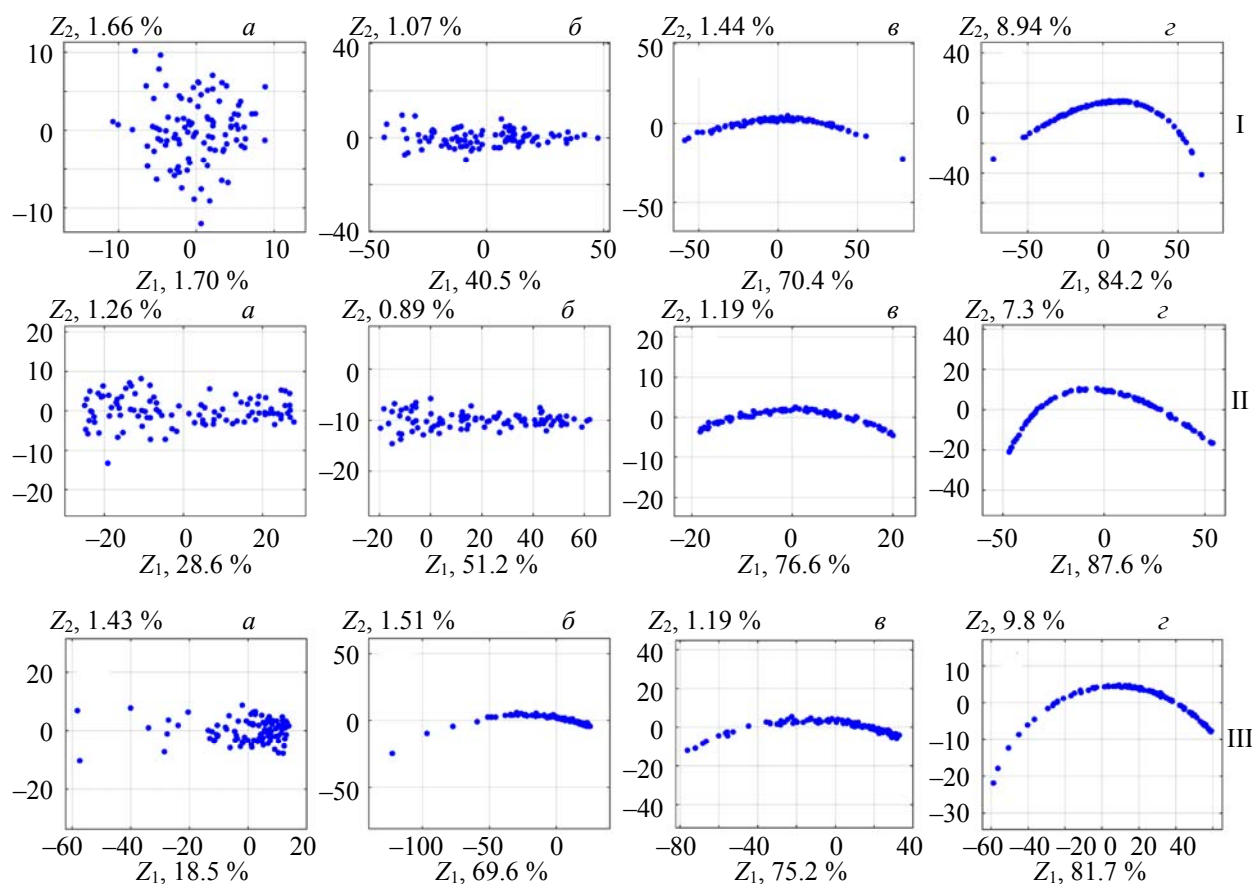


Рис. 3. Диаграммы МГК первых двух главных компонент, полученные в результате анализа наборов данных ГЧФ для системы 3. Распределения времени затухания τ : I — нормальное со средним, равным $\tau = 2$ нс, и $\sigma = 0$ (а), 0.1 (б), 0.2 (в) и 0.5 нс (г); II — равномерное в интервалах: [1.9, 2.1] (а), [1.8, 2.2] (б), [1.6, 2.4] (в) и [1.0, 3.0] нс (г); III — показательное с $\sigma = 0.1$ (а), 0.2 (б), 0.4 (в) и 1.0 нс (г) при $\tau = 2.0$ нс

При увеличении дисперсии параметра моделируемой системы кластеры точек в пространстве ГК вытягиваются в линию, затем выгибаются в дугу. Форма кластера определяется функцией распределения вероятностей параметра моделируемой системы, что особенно заметно в случае показательного распределения.

Для исключения внесения ошибки в интерпретацию причин появления дуги аналогичный численный эксперимент проведен для системы 2, являющейся упрощением системы 1. Результаты применения МГК к набору сгенерированных ГЧФ для системы 2 показаны на рис. 4.

Форма кластеров в пространстве ГК для системы 2 сходна с контурами кластеров системы 3 и определяется в первую очередь не физикой процессов конкретной моделируемой системы, а видом распределения и дисперсией некоторого параметра системы.

Изогнутость и сжатие облака данных в пространстве ГК сильно затрудняет разделение кластеров многомерных данных и, следовательно, их интерпретацию. Для устранения эффекта нелинейности в МГК-пространстве кластеров точек предлагается использовать нормировку локального взвешенного сглаживания (lowess) [15]. Применение этой нормировки для системы 1 позволяет устранить эффект дуги и сжатия кластеров ГЧФ (рис. 5) и таким образом расширяет возможности применения МГК, повышает эффективность последующего анализа и интерпретации кластеров данных.

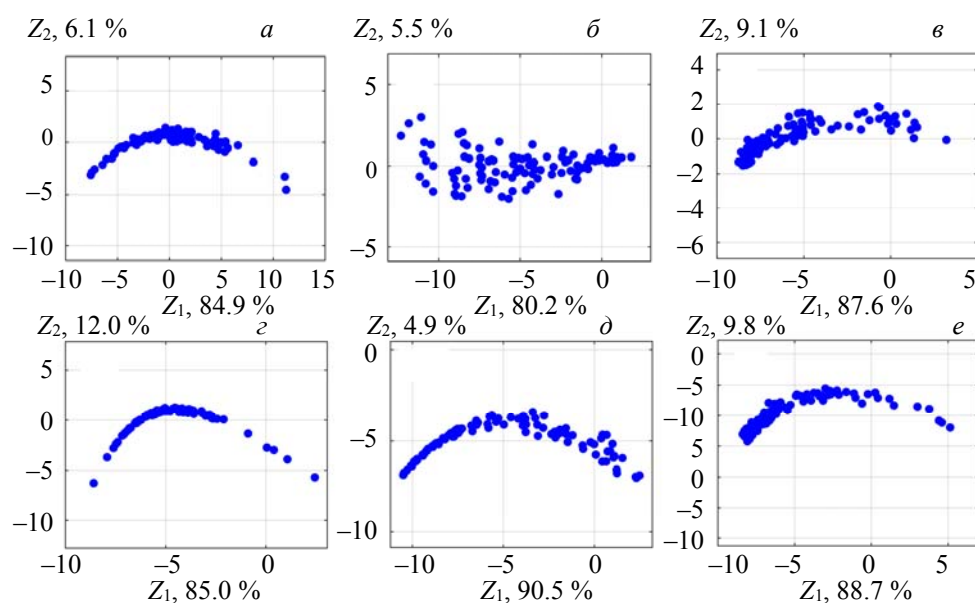


Рис. 4. Диаграммы МГК первых двух главных компонент, полученные в результате анализа наборов данных ГЧФ системы 2. Распределения параметра яркости: нормальное со средними, равными $1 \cdot 10^5$ срms, и $\sigma 2 \cdot 10^4$ (а) и $3 \cdot 10^4$ срms (г); равномерное на отрезках $[8 \cdot 10^4, 1.2 \cdot 10^5]$ (б) и $[5 \cdot 10^4, 1.5 \cdot 10^5]$ срms (д) и показательное с параметрами $2 \cdot 10^4$ (в) и $5 \cdot 10^4$ срms (е)

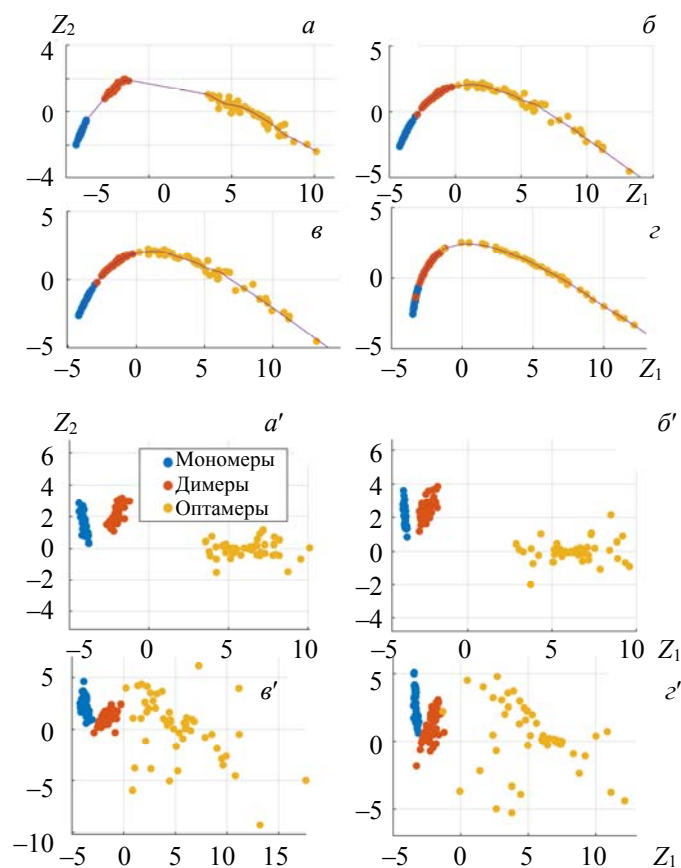


Рис. 5. Диаграммы МГК первых двух главных компонент, полученные до (а—г) и после (а'—г') применения нормировки локального взвешенного сглаживания для системы 1. Исходные данные представлены на рис. 2, б—д

В ходе имитационного моделирования экспериментальной молекулярной системы установлен эффект нелинейности данных в пространстве первых двух ГК, что предполагает наличие значительного разброса по одному или нескольким физическим параметрам. Учитывая, что разброс параметров также является характеристикой системы, полукруглая форма облака данных позволяет предположить дополнительную важную информацию — большую изменчивость параметров. Во флуоресцентной флукуационной спектроскопии, согласно [12], сферообразное облако ожидается для мономеров, димеров и даже агрегатов некоторого белка с относительно стабильными характеристиками, а эллипсоидальное — для смеси молекул нескольких типов. Однако в нашем исследовании вытянутость облака может быть результатом увеличения дисперсии измеряемого параметра, что свидетельствует о формировании агрегатов определенного типа или наличии неустраняемых экспериментальных эффектов (фотообесцвечивания, неконтролируемого лигандирования мономеров неудаляемыми примесями и др.).

Таким образом, для системы, представленной на рис. 1, *а*, причиной нелинейности данных в пространстве ГК, вероятно, является существенный разброс по яркости. Однако, учитывая систему, представленную на рис. 1, *б*, облако-дуга может быть следствием большой вариативности концентрации белка FKBP12-GFP в клетках, дополнительно усиленной процессом фотообесцвечивания (существенным в данном эксперименте). Подтверждением последнего вывода является схожесть рис. 1, *б* и 2, *д* (асимметричная форма дуги с пологим вытянутым хвостом). Следовательно, появление облака-дуги данных свидетельствует о сильной неоднородности системы, что может быть использовано в экспериментальных исследованиях для оценки однородности (качества) подготовленных молекулярных образцов и/или устойчивости процессов, происходящих в них.

Анализ ГЧФ с помощью методов PCH, FIDA, BDGA позволяет оценить физические параметры исследуемой системы. Однако программная реализация часто затруднена, так как модель системы не представлена простой математической формулой, а рассчитывается с использованием сложного численного алгоритма [22], в то время как анализ больших наборов данных занимает много времени. В противоположность этому, МГК является быстрым и простым в программной реализации. Не применяя сложных параметрических методов оценки параметров ГЧФ, по виду облака данных в пространстве первых ГК можно предположить наличие или отсутствие интересных/значимых процессов: сфера — один тип молекул определенной яркости и концентрации, эллипсоид — незначительный разброс в яркости/концентрации, дуга — существенный разброс в яркости/концентрации. Это значительно упрощает принятие решения о целесообразности последующего углубленного анализа, например, с использованием методов PCH, FIDA и BDGA.

Закключение. На основе имитационного моделирования испускания флуоресценции различных молекулярных систем выявлена причина появления дуги в пространстве главных компонент — значительный разброс в некотором параметре(ах) исследуемой системы. Форма дуги определяется видом распределения параметра. В отсутствие разброса в параметрах системы в пространстве главных компонент наблюдается сферообразное облако точек, переходящее в эллипс при небольшом разбросе параметра(ов). При увеличении дисперсии параметра эллипс вытягивается в линию и затем переходит в дугообразное облако. Данный эффект наблюдается при варьировании любого из моделируемых параметров систем и, следовательно, не зависит от физической сути процессов, протекающих в системе, а определяется существенным разбросом в параметрах системы.

Наличие дуги в пространстве главных компонент для экспериментально исследуемых систем является подтверждением сильной гетерогенности системы относительно некоторого параметра(ов) и может использоваться как дополнительный источник информации о системе. Так, при анализе гистограммы числа фотоотсчетов в системах флуоресцентной флукуационной спектроскопии с помощью метода главных компонент наличие дуги может свидетельствовать о существенном разбросе как яркости, так и концентрации молекул. Фотодеструктивные процессы могут также способствовать увеличению разброса в параметрах системы.

Для нормировки результатов метода главных компонент успешно применен метод локального взвешенного сглаживания. Применение нормировки приводит к существенному улучшению идентификации и интерпретации кластеров данных. Достоинство метода главных компонент — использование простого вращения системы координат исходных признаков, или линейного преобразования экспериментальных данных, что не приводит к изменению геометрических расстояний между объектами и, соответственно, повышает эффективность последующего анализа данных.

Предложенное нормирование данных может использоваться в методе комплексного анализа флуктуаций интенсивности флуоресценции молекулярных соединений с применением алгоритмов интеллектуального анализа данных для определения структурного состава олигомеров белков, а также в качестве дополнительных вычислительных процедур к классическим методам анализа РСН, FIDA и BDGA.

- [1] **E. L. Elson, D. Magde.** *Biopolymers*, **13**, N 1 (1974) 1—27, doi: 10.1002/BIP.1974.360130102
- [2] **A. Kitamura, M. Kinjo.** *Int. J. Mol. Sci.*, **19**, N 4 (2018) 964(1—18), doi: 10.3390/IJMS19040964
- [3] **M. A. Hink.** *Protoplasma*, **251**, N 2 (2014) 307—316, doi: 10.1007/S00709-013-0602-Z
- [4] **Y. Chen, J. D. Müller, P. T. C. So, E. Gratton.** *Biophys. J.*, **77**, N 1 (1999) 553—567, doi: 10.1016/S0006-3495(99)76912-2
- [5] **P. Kask, K. Palo, D. Ullmann, K. Gall.** *Proc. Nat. Acad. Sci. USA*, **96**, N 24 (1999) 13756—13761, doi: 10.1073/PNAS.96.24.13756P
- [6] **B. Huang, T. D. Perroud, R. N. Zare.** *Chem. Phys. Chem.*, **5**, N 10 (2004) 1523—1531, doi: 10.1002/CPHC.200400176
- [7] **T. D. Perroud, B. Huang, R. N. Zare.** *Chem. Phys. Chem.*, **6**, N 5 (2005) 905—912, doi: 10.1002/cphc.200400547
- [8] **K. Palo, Ü. Mets, S. Jäger, P. Kask, K. Gall.** *Biophys. J.*, **79**, N 6 (2000) 2858—2866, doi: 10.1016/S0006-3495(00)76523-4
- [9] **K. Palo, Ü. Mets, V. Loorits, P. Kask.** *Biophys. J.*, **90**, N 6 (2006) 2179—2191, doi: 10.1529/BIOPHYSJ.105.066084
- [10] **V. V. Skakun, R. Engel, A. V. Digris, J. W. Borst, A. J. W. G. Visser.** *Front Biosci. (Elite Ed.)*, **3**, N 2 (2011) 489—505, doi: 10.2741/E264
- [11] **L. M. Nederveen-Schippers, P. Pathak, I. Keizer-Gunnink, A. H. Westphal, P. J. M. van Haastert, J. W. Borst, A. Kortholt, V. V. Skakun.** *Int. J. Mol. Sci.*, **22**, N 14, 7300 (2021), doi: 10.3390/IJMS22147300
- [12] **Н. Н. Яцков, В. В. Скакун, Л. Недервин-Шипперс, А. Кортхольт, В. В. Апанасович.** *Журн. прикл. спектр.*, **87**, № 4 (2020) 628—636 [**M. M. Yatskou, V. V. Skakun, L. Nederveen-Schippers, A. Kortholt, V. V. Apanasovich.** *J. Appl. Spectrosc.*, **87**, N 4 (2020) 685—692, doi: 10.1007/S10812-020-01055-6]
- [13] **I. T. Jolliffe, J. Cadima.** *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**, N 2065 (2016), doi: 10.1098/RSTA.2015.0202
- [14] **P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarkegh, P. Gardner.** *Analyst.*, **135**, N 2 (2010) 268—277, doi: 10.1039/B921056C
- [15] **W. S. Cleveland.** *J. Am. Stat. Ass.*, **74**, N 368 (1979) 829—836
- [16] **P. J. M. van Haastert, I. Keizer-Gunnink, A. Kortholt.** *J. Cell Biol.*, **177**, N 5 (2007) 809—816, doi: 10.1083/JCB.200701134
- [17] **Н. Н. Яцков.** *Интеллектуальный анализ данных: пособие*, Минск, БГУ (2014) [**M. M. Yatskou.** *Data Mining: Manual [in Russ.]*, BSU, Minsk (2014)]
- [18] **I. P. Shingaryov, V. V. Skakun, V. V. Apanasovich.** *Methods Mol. Biol.*, **1076** (2014) 743—755, doi: 10.1007/978-1-62703-649-8_34
- [19] **J. R. Lakowicz.** *Principles of Fluorescence Spectroscopy*, 3rd ed., Springer, New York (2006), doi: 10.1007/978-0-387-46312-4
- [20] **M. M. Yatskou.** *Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Systems*, Wageningen (2001)
- [21] **W. S. Cleveland.** *The American Statistician*, **35**, N 1 (1981) 54, doi: 10.2307/2683591
- [22] **V. V. Skakun, A. V. Digris, V. V. Apanasovich.** *Methods Mol. Biol.*, **1076** (2014) 719—741, doi: 10.1007/978-1-62703-649-8_33