МЕТОД ОБРАБОТКИ КИНЕТИЧЕСКИХ КРИВЫХ ЗАТУХАНИЯ ФЛУОРЕСЦЕНЦИИ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Н. Н. Яцков [1*], В. В. Скакун [1], В. В. Апанасович [2]

УДК 535.37

[1] Белорусский государственный университет,
220030, Минск, Беларусь; e-mail: yatskou@bsu.by
[2] Институт информационных технологий и бизнес-администрирования,
220004, Минск, Беларусь

(Поступила 7 декабря 2019)

Предложен метод обработки больших наборов кинетических кривых затухания флуоресценции молекул с использованием алгоритмов интеллектуального анализа данных, позволяющий определить параметры биофизических и оптических процессов, протекающих в молекулярных системах. Идея разработанного метода состоит в разбиении исходного набора кривых затухания флуоресценции на кластеры по степени близости в некоторой мере сходства, нахождении медоидов кластеров, применении метода снижения размерности данных и отображении экспериментальных данных в пространстве двух-, трехмерной размерности, анализе кривых затуханий медоидов с использованием аналитических или имитационных моделей. Применимость метода рассматривается на примере анализа наборов данных, представляющих собой системы флуорофоров. Разработанный метод требует существенно меньше времени и вычислений аналитической функции аппроксимации, а точность оцененных параметров выше, чем в случае применения классического подхода.

Ключевые слова: *флуоресцентная спектроскопия, флуорофор, затухание флуоресценции, имитационное моделирование, интеллектуальный анализ данных.*

A method for processing big datasets of the kinetic curves of fluorescence decay using data mining algorithms is proposed to determine the parameters of biophysical and optical processes that occur in molecular systems. The idea of the developed method is in partitioning the initial fluorescence dataset into clusters according to the degree of likeness to some measure of similarity, finding cluster medoids, using a data reduction method and visualizing experimental data in two- or three-dimensional space, analyzing the fluorescence curves of the medoids by analytical or simulation models. The applicability of the method is considered by the example of the analysis of datasets representing systems of fluorophores. The developed method uses substantially less time and computation of the analytical approximation function, though the accuracy of the estimated parameters is higher than in the classical approach.

Keywords: *fluorescence spectroscopy, fluorophores, fluorescence decay, simulation modelling, data mining.*

Введение. Методы экспериментальной флуоресцентной спектроскопии и микроскопии интенсивно применяются в физико-химических, биофизических и биомедицинских исследованиях для определения локализации, динамики и взаимодействия молекулярных соединений, а также параметров оптических процессов, протекающих в сложных системах [1—3]. Современные методы экспериментальных исследований, такие как однофотонного счета (time-correlated single photon counting), фазово-модуляционный (phase-shift fluorimetry), микроскопии визуализации времени жизни флуорес-

---

METHOD FOR PROCESSING THE KINETIC CURVES OF FLUORESCENCE DECAY USING DATA MINING ALGORITHMS

M. M. Yatskou [1*], V. V. Skakun [1], V. V. Apanasovich [2] (*[1] Belarusian State University, Minsk, 220030, Belarus; [2] Institute of Information Technology and Business Administration, Minsk, 220004, Belarus; e-mail: yatskou@bsu.by*)

ценции (fluorescence lifetime imaging microscopy, FLIM), позволяют регистрировать большие наборы кинетических кривых затухания флуоресценции молекулярных систем в различных окружениях [4—7].

Для анализа кинетических кривых затухания флуоресценции применяются методы оптимизации, различные аналитические и имитационные модели оптических процессов, протекающих в молекулярных системах [8—11]. Классический анализ большого набора кривых затуханий флуоресценции состоит в отдельной обработке каждого набора данных, что имеет ряд ограничений: необходимость значительных временных и вычислительных ресурсов для выполнения процедур поиска параметров моделей, что особенно критично при использования имитационных моделей; сложность выбора начальных приближений в процедурах оценки параметров; потеря точности в оценке параметров моделей в условиях высокого экспериментального шума, что характеризуется появлением локальных минимумов на поверхности целевой функции. Полным или частичным устранением вышеперечисленных ограничений является применение алгоритмов интеллектуального анализа, использующих парадигму одновременного анализа всего набора данных как единого целого [12—16].

В настоящей работе предложен метод обработки больших наборов кинетических кривых затухания флуоресценции молекул с использованием алгоритмов интеллектуального анализа данных с целью определения параметров биофизических и оптических процессов в молекулярных системах. Разработанный метод исследуется на имитационно смоделированных наборах данных микроскопии визуализации времени жизни флуоресценции, представляющих собой три системы флуорофоров и характеризующихся одно-, двух- и стрэтч-экспонециальным законами испускания флуоресценции.

Методология. Идея разработанного метода анализа состоит в разбиении исходного набора кривых затухания флуоресценции на кластеры по степени близости в заданной мере сходства, нахождении медоидов кластеров, применении метода снижения размерности данных и отображении экспериментальных данных в пространстве двух-, трехмерной размерности, анализе кривых затуханий медоидов с использованием аналитических или имитационных моделей. Схема метода представлена на рис. 1. Рассмотрим основные этапы метода.

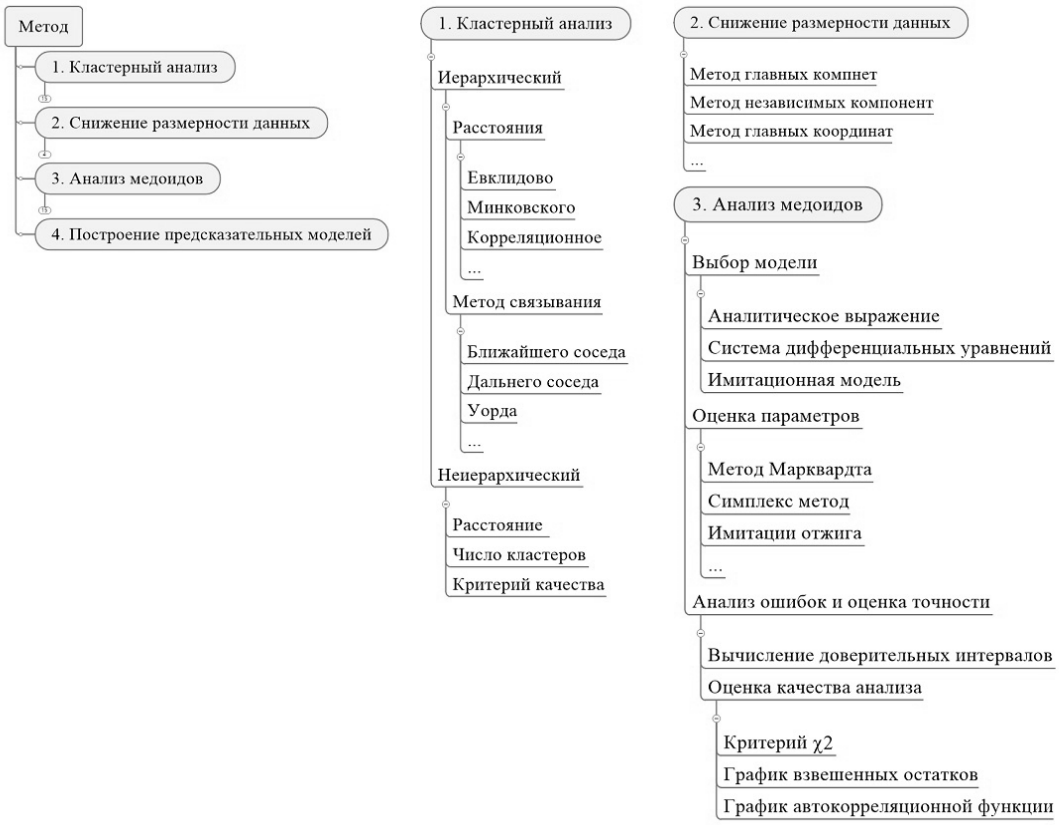

Рис. 1. Блок-схема метода и его основных этапов исследования флуоресценции молекулярных соединений с использованием алгоритмов интеллектуального анализа данных

Кластерный анализ. Объединение схожих кривых затухания флуоресценции является задачей кластерного анализа. Кластерный анализ не требует априорной информации о целевых метках классов данных и позволяет разделить множество исследуемых объектов на группы похожих объектов — кластеры. По способам кластеризации методы кластерного анализа можно условно разделить на две большие группы — иерархические и неиерархические методы [12, 17]. Каждая из групп методов включает множество подходов и алгоритмов. Для кластерного анализа небольших наборов данных (<5000 объектов) предпочтительнее использовать медленные и более точные иерархические методы, для анализа больших наборов данных (>5000 объектов) — быстрые и менее точные методы неиерархического анализа. Для нахождения определенного решения задачи иерархического кластерного анализа необходимо задать способ сравнения объектов между собой (меру сходства, например евклидово, Минковского, корреляционное расстояния), способ кластеризации (ближнего или дальнего соседа, Уорда), разбиение данных по кластерам (установление числа кластеров). Критерием для определения схожести объектов может быть некоторое расстояние между точками объектов на диаграмме рассеяния. В неиерархических методах кластерного анализа необходимо задать гипотезу о числе кластеров k, способ сравнения объектов между собой (меру сходства или расстояние) и функцию качества или критерия кластеризации. В основе неиерархических методов лежит разбиение набора данных на заданное число кластеров таким образом, чтобы целевая функция алгоритма (функция качества кластеризации) достигала экстремума.

Выполняется кластерный анализ кривых затухания флуоресценции в пространстве исходных признаков, представленных числом фотоотсчетов во временных каналах гистограмм регистрируемых фотонов. Определяются кластеры кривых затуханий флуоресценции в некоторой мере сходства. Вычисляются медоиды кластеров — кривые затухания, имеющие наименьшие средние расстояния до остальных объектов соответствующих кластеров.

Снижение размерности данных. Учет большой группы малоинформативных признаков, представленных количеством фотоотсчетов в заданные моменты времени, приводит к затруднению анализа данных, а именно их зашумлению, увеличению объема данных, искажению достоверной информации о кластерах схожих кривых затуханий. Для улучшения качества анализа данных, в частности визуальной оценки разбиения данных на кластеры, перспективным направлением является проведение этапа анализа данных, включающего в себя переход в пространство невысокой размерности новых информативных признаков кривых затухания флуоресценции, в котором кривые затухания флуоресценции формируют кластеры. Для выполнения данного преобразования требуется применение алгоритмов снижения размерности данных [12, 18]. Снижение размерности данных — наиболее эффективный подход для удаления шумовых и неинформативных атрибутов объектов. Методы снижения размерности данных включают в себя две группы алгоритмов: на основе построения (не)линейных комбинаций признаков и на основе выделения наиболее информативных исходных признаков. К алгоритмам первой группы относятся методы главных и независимых компонент, главных координат, дискриминантный и факторный анализ [18—20]. Идея алгоритмов данной группы состоит в переходе в пространство низкой размерности (новых признаков) без потери сущности информации. Идея алгоритмов второй группы состоит в выделении небольшой группы исходных наиболее информативных признаков объектов, минимизирующих шум и избыточность в данных и максимизирующих их информативность в смысле разделения на кластеры или классы [21, 22]. Часто применение алгоритмов второй группы предпочтительнее, так как не приводит к изменению исходных данных, в то время как в пространстве новых признаков взаимное расположение кластеров данных может измениться, что приведет к неверной биофизической интерпретации исследуемых процессов. В данной работе в первую очередь требуется сокращение пространства признаков кривых затухания флуоресценции, поэтому используется метод линейного преобразования признаков, а именно метод главных компонент (МГК). МГК является базовым алгоритмом сжатия данных и интенсивно используется в ходе анализа ИК-фурье-спектров [23, 24]. Он обладает высоким быстродействием, что обеспечивает проведение вычислительного эксперимента при малых вычислительных затратах, относительно простым для программной реализации.

Применяется МГК к набору кривых затуханий флуоресценции. Отображаются кластеры и их медоиды на диаграмме рассеяния первых главных компонент. Если кластеры данных не разделяются, то можно полагать, что присутствует только один вид молекулярных соединений. Иначе допускается наличие нескольких форм молекулярных соединений (флуорофоров).

Анализ медоидов. Для аппроксимации кинетических кривых затухания флуоресценции, представленных найденными медоидами кластеров, традиционно применяются системы дифференциальных уравнений, описывающие пространственно-временные состояния молекул, многоэкспоненциальные модели, различные аналитические и имитационные модели описания фотофизических процессов [8, 9]. Для оптимального подбора параметров математических моделей в ходе аппроксимации экспериментальных данных используются методы оптимизации Левенберга—Маргквардта, Нелдера—Мида, имитации отжига [9]. Наилучшая аппроксимация определяется критерием (или набором критериев), определяющим степень отклонения теоретической функции от экспериментальных данных. Как правило, такой критерий представляется аналитически в виде функции экспериментальных и смоделированных данных, вид которой определяется областью применения, непосредственным методом моделирования и условиями проведения эксперимента. В качестве критериев качества широко используются как количественные критерии χ^2, Колмогорова—Смирнова, Романовского, так и диаграммы — графики взвешенных остатков, их автокорреляционной функции и гистограммы. Для оцененных параметров математических моделей строятся доверительные интервалы с помощью алгоритмов методов Монте-Карло, асимптотических стандартных ошибок, исчерпывающего поиска (exhaustive search) [9]. Для точного определения параметров молекулярных соединений выполняется анализ медоидов каждого кластера с использованием классических алгоритмов оптимизации и математических моделей.

Построение предсказательных моделей. Интерпретация — объяснение и улучшение понимания объектов системы и их поведения. В ходе данного этапа исследуется поведение элементов системы на каких-либо критических точках, для которых получение данных недоступно по техническим или экономическим причинам. Решаются задачи оптимизации и оценки чувствительности системы. Задача оптимизации — точное определение такого сочетания факторов, параметров и их величин, при котором обеспечивается наилучший показатель качества системы. Анализ чувствительности — выявление факторов, в наибольшей степени влияющих на функционирование системы. Прогнозирование представляет собой предсказание поведения исследуемой системы на основе разработанной модели. Прогнозирование — главная цель моделирования и оценки поведения молекулярной системы при некотором сочетании ее управляемых и неуправляемых параметров [9, 12].

Имитационное моделирование. Рассмотрены смоделированные данные, позволяющие качественно и количественно оценить работоспособность разработанного и классического методов анализа данных. Реализованы имитационные модели кинетических кривых затуханий флуоресценции трех систем молекул в некотором абстрактном FLIM эксперименте (системы 1—3): двух люминесцирующих молекулярных мономеров, пространственно разделенных на некоторой поверхности; двух люминесцирующих молекулярных мономеров, частично смешанных на поверхности; люминесцирующего молекулярного донора, окруженного нелюминесцирующими молекулами акцептора.

Система 1 описывается одноэкспоненциальной моделью затухания. Экспоненциальная модель является наиболее простой моделью интенсивности затухания флуоресценции, используется для математического описания растворов низких концентраций невзаимодействующих молекул [1], интенсивность которых

$$I(t, I_0, \tau) = I_0 e^{-t/\tau}, \tag{1}$$

где τ и I_0 — время затухания и интенсивность флуоресценции в момент времени $t = 0$ (относительно импульса возбуждения). Моделирование момента времени t_{det} прибытия фотона на детектор проводится по формуле [9]

$$t_{\text{det}} = -\tau \ln(z), \tag{2}$$

где z — реализация случайной величины, равномерно распределенной в диапазоне [0,1]. Разыгрываются возбуждения одной молекулы, регистрируется момент времени t_{det} и заносится в канал гистограммы $K_{\text{det}} = (t_{\text{det}}/\Delta t)$, где Δt — ширина временного канала многоканального анализатора. Времена затухания, представленные временем прибытия фотона t_{det}, регистрируются в гистограммах $i(t,\mathbf{a})$, где $\mathbf{a}$ — вектор параметров модели. Моделирование продолжается до тех пор, пока не накопится заданное число отсчетов в максимуме. Гистограмма времен регистрации фотонов представляет собой затухание интенсивности флуоресценции образца. Измеряемая интенсивность затухания $I(t,\mathbf{a})$ является сверткой функции отклика образца $i(t,\mathbf{a})$ и функции вспышки лазера, обычно представляемой конечной функцией отклика аппаратуры $e(t)$. Математически свертка для интенсивности затухания $I(t,\mathbf{a})$ записывается в виде

$$I(t,a) = e(t) \otimes i(t,a) = \int_0^t e(t-x) i(x,a) dx\,, \tag{3}$$

где $e(t)$ — функция вспышки лазера, моделируется в виде гауссообразного импульса, в простейшем случае — прямоугольного импульса.

Система 2 является частным (двухэкспоненциальным) случаем многоэкспоненциальной модели и описывает раствор невзаимодействующих молекул нескольких видов с временами затухания флуоресценции τ_l, $l = 1, 2, \ldots, L$ [1]. Аналитическое выражение закона затухания многоэкспоненциальной модели имеет вид

$$I(t, p_1, p_2, ..., p_L, \tau_1, \tau_2, ..., \tau_L) = \sum_{l=1}^{L} \frac{p_l}{\tau_l} e^{-t/\tau_l}\,, \tag{4}$$

где $\sum_{l=1}^{L} p_l = 1$.

Для генерации момента регистрации фотона t_{det} на l-й молекуле необходимо разыграть два выборочных значения равномерно распределенной на интервале [0,1] случайной величины z_1 и z_2. Затем интервал [0,1] делится на l каналов, где каждый l-й канал пропорционален вкладу l-го типа молекул [9]. Если z_1 попадает в l-й канал, момент времени t_{det} генерируется по формуле

$$t_{\text{det}} = -\tau_l \ln z_2. \tag{5}$$

Заполнение гистограммы проводится аналогично предыдущему алгоритму. Таким образом, в результате многократных возбуждений системы многоэкспоненциальная кривая затухания флуоресценции регистрируется в гистограмму $i(t,\mathbf{a})$. Выполнение операции свертки формирует интенсивность затухания $I(t,\mathbf{a})$.

Система 3. Модель стрэтч, или "растянутой", экспоненты представляет собой аналитическое описание кинетики затухания интенсивности флуоресценции доноров в донорно-акцепторной системе в присутствии переноса энергии электронного возбуждения по Фёрстеру [1]. Выражение для флуоресценции донора в трехмерном пространстве может быть представлено в виде

$$I(t, I_0, q, \tau_D) = I_0 \exp\{-t/\tau_D - q(t/\tau_D)^{1/2}\}, \tag{6}$$

где I_0 — интенсивность флуоресценции в момент времени $t = 0$; $q = 0.5[C_A]/[C_{A0}]$, C_{A0} и C_A — критическая и естественная концентрация акцепторов, τ_D — время затухания флуоресценции доноров. Имитационная модель для моделирования флуоресценции донора в присутствии переноса энергии по Фёрстеру реализована на основе метода Неймана [25]. В соответствии с данным алгоритмом время затухания донора t_{det} генерируется по формуле (2) и принимается, если $z_1 < \exp\{-t_{\text{det}}/\tau_D - q(t_{\text{det}}/\tau_D)^{1/2}\}$, где z_1 — реализация равномерно распределенной случайной величины на интервале [0,1]. Флуоресценция донора, представленная временем затухания t_{det} и характеризуемая двумя параметрами q и τ_D, регистрируется в гистограмме $i(t,\mathbf{a})$. Вычисляется интенсивность затухания $I(t,\mathbf{a})$.

На рис. 2 представлены примеры наборов данных смоделированных и теоретических кривых затухания флуоресценции для систем 1—3.

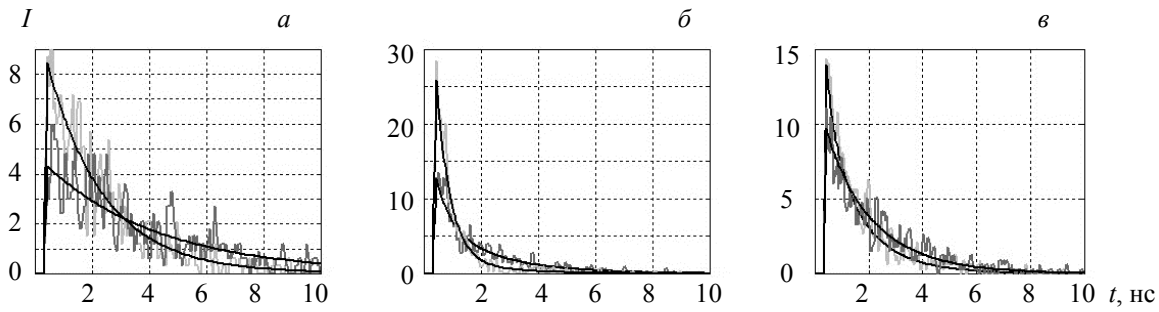

Рис. 2. Примеры наборов данных смоделированных (в градациях серого цвета) и теоретических (черный цвет) кривых затухания флуоресценции для систем 1—3: *а* — система 1, параметры моделирования: $\tau_1 = 2$ и $\tau_2 = 4$ нс; *б* — система 2, параметры моделирования: $\tau_1 = 2$ и $\tau_2 = 4$ нс, их вклады $p_1 = 0.2$, $p_2 = 0.8$ и $p_1 = 0.8$, $p_2 = 0.2$ для двух наборов кривых затуханий; *в* — система 3, параметры моделирования: $\tau_D = 2$ нс, концентрация акцепторов $q = 1$ и 0.2 для двух наборов кривых затуханий

Для учета эффекта разброса в данных или “размытости” кластеров кривых затухания флуоресценции, обусловленных влиянием различных искажений, таких как наличие неустранимых примесей, тушащих или стимулирующих флуоресценцию молекул, высокий фоновый шум, засветка и деградация красителей, используется моделирование параметров моделей, имеющих нормальное распределение с заданным математическим ожиданием a_j и среднеквадратическим отклонением σ (СКО). Варьирование СКО позволяет контролировать разброс в данных или размытость кластеров кривых затуханий в многомерном пространстве временных отсчетов.

Классический метод анализа состоит в отдельной аппроксимации каждой кривой затухания флуоресценции одно-, двух- и стрэтч-экспоненциальными моделями и оценке времен жизни флуорофоров по набору восстановленных времен затуханий отдельных кривых (проводится аппроксимация гистограммы оцененных времен затуханий с использованием модели гауссовых распределений). Более сложные методы, такие как глобальный или таргетный анализ [26, 27], могут быть рассмотрены, однако не включены в данную работу по причине их высокой вычислительной сложности реализации при сопоставимой интерпретируемости результатов.

Описание вычислительного эксперимента. Смоделированы три системы люминесцирующих флуорофоров. Параметры моделирования: длина интервала наблюдения 10 нс, число временных каналов гистограммы 256, количество кривых 200 и 2000, уровень экспериментального шума от 5—10 до 90—110 фотоотсчетов в максимуме, СКО для моделирования разброса в данных $\sigma = 0.1$ и 0.2 от абсолютных значений моделируемых параметров, функция вспышки лазера моделируется в виде прямоугольного импульса $\sim 10^{-2}$ интервала наблюдения. Параметры моделирования систем флуорофоров выбирались таким образом, чтобы максимально исследовать разработанный метод и пределы его применимости. Рассмотрены следующие параметры имитационных моделей систем 1—3.

Система 1. $\tau_1 = 2$ и $\tau_2 = 4$ нс; $\tau_1 = 1.4$ и $\tau_2 = 2$ нс. Пример простой “эталонной” системы, где времена жизни флуоресценции флуорофоров существенно разделяются или достаточно близки, как в случае наличия невзаимодействующих мономеров порфирина в растворе [28]. Используется для проверки принципиальной применимости разработанного метода. Исследуются точность оценки времен жизни и время вычислений в зависимости от числа кривых и разброса/размытости кластеров. Разработанный метод наиболее тщательно исследуется на данной модели, что обусловлено ограничениями имеющихся вычислительных ресурсов.

Система 2. $\tau_1 = 0.5$ нс, $\tau_2 = 2$ нс, $p_1 = 0.2$, $p_2 = 0.8$; $\tau_1 = 0.5$ нс, $\tau_2 = 2$ нс, $p_1 = 0.8$, $p_2 = 0.2$. Пример сложной (в смысле близости времен затухания и многопараметричности) модели молекулярной системы, например, тетрамер и мономер порфирина [10]. Исследуются точность оценки параметров и время вычислений в зависимости от разброса/размытости кластеров.

Система 3. $\tau_D = 2$ нс, $q_A = 1$; $\tau_D = 2$ нс, $q_A = 0.2$. Пример наиболее сложной из рассматриваемых моделей молекулярных систем для поиска глобального минимума целевой функции, что обусловлено вытянутостью и оврагоподобностью изолиний целевой функции, и в условиях высокого экспериментального шума представляет собой пример крайне сложной задачи оптимизации для поиска параметров многомодальной целевой функции [29]. Исследуются точность оценки параметров и время вычислений.

Вычислительные процедуры алгоритмов разработанного и классического методов реализованы в среде математического программирования Matlab с использованием функций pdist, linkage, cluster, eigen, nlinfit, nlparfci, интегрирующих алгоритмы иерархического кластерного анализа, МГК, оптимизации Левенберга—Марквардта и вычисления 90 % доверительных интервалов [30]. Использован иерархический метод кластерного анализа, исследованы наиболее распространенные способы вычисления расстояния — Евклида, Минковского и Кендэла (корреляционное). В МГК применяется процедура центрирования данных. Начальные приближения параметров моделей аппроксимации генерировались случайным образом в области параметров [0;50] или [0;25] в случае сложной сходимости классического метода, характеризующей удаленность начальной точки поиска от положения глобального минимума в пространстве параметров, и выражаются в процентах (%) от абсолютных значений моделируемых параметров. С целью определения наиболее оптимального числа медоидов разработанный метод исследован при использовании 1, 5 и 10 медоидов. Для оценки ошибки точности восстановления параметров рассмотрено среднее отклонение восстановленных параметров от значений, используемых в имитационных моделях:

$$\varepsilon = \frac{1}{n}\sum_{j=1}^{n}\frac{|a_j - a_j^*|}{a_j^*}\cdot 100\%\,, \qquad (7)$$

где a_j^* и a_j — смоделированные и оцененные параметры; n — число параметров имитационной модели. Вычисления проведены на ПК с основными характеристиками DualCore Intel Pentium E5700, 3000 MHz, 8156 MB DDR3-1333 RAM. Операционная система MS Windows 10 Pro, 64 bit.

Результаты и их обсуждение. Результаты анализа смоделированных данных для системы 1 представлены в табл. 1, пример — на рис. 3, *а—г*. Относительная доля разброса, приходящаяся на первые две главные компоненты, 45.4 %, на первые 70 — 95 %. В пространстве первых главных компонент кривые затухания образуют два непересекающихся кластера, что подтверждает наличие двух видов флуорофоров в исследуемой системе.

Т а б л и ц а 1. Параметры и их 90%-ные доверительные интервалы (в квадратных скобках), полученные в результате анализа смоделированных кривых затухания флуоресценции системы 1 с использованием классического и разработанного (на основе 1, 5 и 10 медоидов) методов

Параметры моделирования				Классический метод				Разработанный метод											
								1 медоид				5 медоидов				10 медоидов			
N	σ	τ_1*	τ_2*	τ_1	τ_2	ε,%	t, с	τ_1	τ_2	ε,%	t, с	τ_1	τ_2	ε,%	t, с	τ_1	τ_2	ε,%	t, с
200	0.1	2.00	4.00	2.08	4.09	**3.13**	1.8	2.04	4.06	**1.75**	1.3	2.06	3.85	**3.38**	1.2	2.05	3.92	**2.25**	1.4
				[2.05;2.11]	[4.01;4.18]			[1.98;2.10]	[3.90;4.22]			[2.03;2.09]	[3.77;4.93]			[2.03;2.07]	[3.86;3.98]		
2000	0.1	2.00	4.00	2.06	4.03	**1.88**	11.4	2.16	3.93	**4.88**	11.4	2.03	3.94	**1.50**	12.2	2.01	3.91	**1.37**	12.3
				[2.05;2.07]	[4.00;4.05]			[2.10;2.22]	[3.76;4.09]			[2.01;2.06]	[3.87;4.02]			[1.99;2.03]	[3.85;3.96]		
200	0.2	2.00	4.00	2.09	4.15	**4.13**	1.6	2.18	3.88	**6.00**	1.2	1.98	3.98	**0.75**	1.4	2.02	4.17	**2.63**	1.3
				[2.02;2.15]	[3.96;4.35]			[2.11;2.25]	[3.71;4.04]			[1.95;2.00]	[3.90;4.06]			[2.00;2.03]	[4.11;4.24]		
2000	0.2	2.00	4.00	2.07	4.02	**2.00**	11.5	2.11	3.79	**5.38**	11.4	2.01	3.83	**2.37**	12.1	1.96	3.78	**3.75**	12.3
				[2.05;2.10]	[3.95;4.10]			[2.04;2.17]	[3.64;3.93]			[1.98;2.03]	[3.76;3.90]			[1.94;1.98]	[3.73;3.83]		
200	0.1	1.40	2.00	1.48	2.10	**5.36**	1.6	1.43	2.04	**2.07**	1.4	1.45	2.04	**2.79**	1.3	1.48	2.03	**3.61**	1.4
				[1.46;1.51]	[2.04;2.14]			[1.40;1.46]	[1.98;2.10]			[1.43;1.47]	[2.01;2.07]			[1.47;1.5]	[2.01;2.06]		
2000	0.1	1.40	2.00	1.46	2.06	**3.64**	13.0	1.39	1.90	**2.86**	11.7	1.39	1.92	**2.36**	11.8	1.37	1.94	**2.57**	11.8
				[1.46;1.47]	[2.05;2.06]			[1.36;1.42]	[1.84;1.95]			[1.37;1.40]	[1.89;1.94]			[1.36;1.38]	[1.92;1.96]		
200	0.2	1.40	2.00	1.66	2.69	**26.54**	1.6	1.22	1.96	**7.43**	1.2	1.22	1.91	**8.68**	1.3	1.21	1.92	**8.79**	1.4
				[1.62;1.69]	[2.49;2.89]			[1.19;1.24]	[1.89;2.02]			[1.21;1.23]	[1.88;1.94]			[1.20;1.22]	[1.89;1.94]		
2000	0.2	1.40	2.00	1.40	1.87	**3.25**	11.8	1.43	2.19	**5.82**	11.6	1.41	2.23	**6.11**	11.8	1.42	2.22	**6.21**	11.2
				[1.39;1.42]	[1.82;1.92]			[1.40;1.46]	[2.13;2.26]			[1.40;1.43]	[2.20;2.26]			[1.41;1.43]	[2.19;2.24]		

Сравниваемые методы успешно справились с нахождением параметров исследуемых молекулярных систем. Время вычислений методов сопоставимо, однако количество вычислений целевой функции в классическом методе на несколько порядков превышало количество вычислений в разработанном методе. В случае небольшого числа кривых затуханий флуоресценции и большого разброса данных (σ = 0.2) разработанный метод (ε = 0.75 и 7.43 %) превосходит по точности оценки параметров кривых затуханий классический подход (ε = 4.13 и 26.54 %). Следует отметить, что точность классического метода может быть существенно повышена при использовании глобального анализа, однако вычислительная сложность решаемой задачи значительно повысится. При увеличении числа кривых точность методов повышается, что обусловлено увеличением выборки обрабатываемых данных и, соответственно, повышением статистической мощности анализа данных. Рассмотрение 5 и 10 медоидов не приводит к существенному увеличению эффективности разработанного алгоритма.

Результаты анализа смоделированных данных для системы 2 представлены в табл. 2, пример — на рис. 3, *д—з*. Относительная доля разброса, приходящаяся на первые две главные компоненты, 77.7 %, на первые 25 — 95 %. Увеличение относительного разброса по сравнению с системой 1 обусловлено усилением вытянутости (элипсоиадальности) облака данных в пространстве признаков кривых затуханий для рассматриваемых параметров модели. В пространстве первых главных компонент кривые затухания образуют два непересекающихся кластера, что подтверждает наличие двух семейств кривых затухания.

Т а б л и ц а 2. Параметры и их 90%-ные доверительные интервалы (в квадратных скобках), полученные в результате анализа смоделированных кривых затухания флуоресценции системы 2 с использованием классического и разработанного (на основе 1, 5 и 10 медоидов) методов

Параметры моделирования						Классический метод							
N	σ	p_1*	p_2*	τ_1*	τ_2*	p_1	p_2	τ_1	τ_2	ε,%	t, с		
200	0.1	0.20	0.80	0.50	2.00	0.21	0.85	0.53	2.17	**6.44**	3.9		
						[0.17;0.25]	[0.82; 0.88]	[0.52; 0.54]	[2.11; 2.23]				
200	0.2	0.20	0.80	0.50	2.00	0.23	0.85	0.50	2.02	**5.56**	4.1		
						[0.19;0.27]	[0.77; 0.93]	[0.49; 0.51]	[1.95; 2.09]				
Разработанный метод													
N	σ	p_1*	p_2*	τ_1*	τ_2*	1 медоид							
						p_1	p_2	τ_1^1	τ_2^1	τ_1^2	τ_2^2	ε,%	t, с
200	0.1	0.20	0.80	0.50	2.00	0.18	0.87	0.39	2.05	0.56	2.43	**12.79**	1.2
						[0.12; 0.24]	[0.78; 0.95]	[0.29; 0.48]	[1.85; 2.26]	[0.52; 0.60]	[0.67; 4.2]		
200	0.2	0.20	0.80	0.50	2.00	0.12	0.84	0.32	1.79	0.49	3.87	**31.17**	1.3
						[0.08; 0.17]	[0.81; 0.87]	[0.23; 0.41]	[1.66; 1.92]	[0.48; 0.51]	[2.35; 5.4]		
N	σ	p_1*	p_2*	τ_1*	τ_2*	5 медоидов							
						p_1	p_2	τ_1^1	τ_2^1	τ_1^2	τ_2^2	ε,%	t, с
200	0.2	0.20	0.80	0.50	2.00	0.17	0.78	0.39	1.97	0.48	2.01	**7.58**	1.3
						[0.13; 0.20]	[0.74; 0.81]	[0.34; 0.44]	[1.88; 2.07]	[0.47; 0.5]	[1.66; 2.35]		
200	0.2	0.20	0.80	0.50	2.00	0.28	0.78	0.57	2.18	0.49	2.40	**14.5**	1.4
						[0.23; 0.33]	[0.75; 0.81]	[0.51; 0.64]	[2.03; 2.33]	[0.48; 0.50]	[2.03; 2.78]		
N	σ	p_1*	p_2*	τ_1*	τ_2*	10 медоидов							
						p_1	p_2	τ_1^1	τ_2^1	τ_1^2	τ_2^2	ε,%	t, с
200	0.2	0.20	0.80	0.50	2.00	0.17	0.74	0.52	1.97	0.49	1.98	**5.17**	1.5
						[0.14; 0.20]	[0.71; 0.77]	[0.46; 0.57]	[1.90; 2.04]	[0.47; 0.5]	[1.74; 2.24]		
200	0.2	0.20	0.80	0.50	2.00	0.21	0.73	0.48	2.02	0.46	2.13	**5.54**	1.4
						[0.17; 0.24]	[0.71; 0.76]	[0.43; 0.53]	[1.94; 2.11]	[0.45; 0.47]	[1.90; 2.36]		

П р и м е ч а н и е. Верхние индексы 1, 2 — номера кластеров.

Применение алгоритмов разработанного метода к анализу набора смоделированных данных позволило точно определить времена жизни флуоресценции флуорофоров. Точность оцененных параметров разработанным методом выше, чем в результате применения классического подхода, успешная сходимость которого достигалась только в случае выбора начальных приближений в области [0;25 %] действительных значений параметров. Разработанный метод требует существенно меньше времени и количества вычислений аналитической функции аппроксимации. Наилучшие результаты получены при рассмотрении 10 медоидов.

Результаты анализа смоделированных данных для системы 3 представлены в табл. 3 и на рис. 3, *и*—*м*. Относительная доля разброса, приходящаяся на первые две главные компоненты, 42.6 %, на первые 70 — 95 %. В пространстве первых главных компонент кривые затухания образуют два частично пересекающихся кластера, что свидетельствует о сложности разделения кривых затухания различных видов. Однако алгоритм иерархического анализа успешно решает задачу точного нахождения кластеров кривых затухания различных видов.

Применение классического метода к анализу набора смоделированных данных не позволило точно определить параметры модели стрэтч-экспоненты (ε = 57.33 %), в то время как разработанный справился удовлетворительно (ε = 9.13 %). Разработанный метод требует меньше времени и количества вычислений аналитической функции аппроксимации. Наилучшие результаты получены при использовании 5 и 10 медоидов.

Относительная доля разброса по первой главной компоненте существенно превышает значения по остальным компонентам в рассматриваемых системах (табл. 4), что позволяет предположить наличие доминирующего фактора в исследуемых физических процессах, вносящего наибольший

вклад в суммарную дисперсию [19]. Первая главная компонента является наиболее значимым или информативным признаком в смысле разделения кластеров [22], что подтверждает хорошую кластеризацию в пространстве первых двух главных компонент при относительно невысоких дисперсиях данных, соответствующих этим компонентам. Диаграмма рассеяния в пространстве неинформативных второй и третьей главных компонент для системы 1 (аналогично выглядят диаграммы для систем 2 и 3) представлена на рис. 4, *а*.

Т а б л и ц а 3. Параметры и их 90%-ные доверительные интервалы (в квадратных скобках), полученные в результате анализа смоделированных кривых затухания флуоресценции системы 3 с использованием классического и разработанного (на основе 1, 5 и 10 медоидов) методов

Параметры моделирования					Классический метод					
N	σ	q^*_1	q^*_2	τ^*_D	q_1	q_2	τ_D	ε,%	t, с	
200	0.1	1.0	0.2	2.00	0.48	–0.03	1.90	**57.33**	2.2	
					[0.38; 0.59]	[–0.11; 0.06]		[1.86; 1.94]		
					Разработанный метод					
N	σ	q^*_1	q^*_2	τ^*_D	1 медоид					
					q_1	q_2	τ_D^1	τ_D^2	ε,%	t, с
200	0.1	1.0	0.2	2.00	1.08	0.10	1.82	2.05	**17.38**	1.4
					[0.83; 1.33]	[–0.1; 0.29]	[1.58; 2.05]		[1.74; 2.36]	
N	σ	q^*_1	q^*_2	τ^*_D	5 медоидов					
					q_1	q_2	τ_D^1	τ_D^2	ε,%	t, с
200	0.1	1.0	0.2	2.00	0.94	0.15	1.91	1.98	**9.13**	1.4
					[0.82; 1.06]	[0.08; 0.23]	[1.82; 2.00]		[1.83; 2.13]	
N	σ	q^*_1	q^*_2	τ^*_D	10 медоидов					
					q_1	q_2	τ_D^1	τ_D^2	ε,%	t, с
200	0.1	1.0	0.2	2.00	0.91	0.24	2.09	1.92	**9.38**	1.8
					[0.83; 1.00]	[0.16; 0.32]	[1.98; 2.20]		[1.83; 2.02]	

П р и м е ч а н и е. Верхние индексы 1, 2 — номера кластеров.

Т а б л и ц а 4. Относительная доля разброса (%), приходящаяся на первые 10 главных компонент, полученная в результате анализа систем 1—3 с использованием метода главных компонент

Система	1	2	3	4	5	6	7	8	9	10
1	**42.5**	**2.9**	2.6	2.4	2.3	2.1	2.0	1.8	1.8	1.7
2	**72.9**	**4.8**	2.2	1.9	1.4	1.3	1.3	1.0	0.9	0.8
3	**38.3**	**4.3**	3.7	3.1	3.0	2.6	2.4	2.2	2.0	1.9

П р и м е ч а н и е. Параметры моделирования указаны в подписи к рис. 3.

Оцененная точность классического и разработанного (для 1, 5 и 10 медоидов) методов для смоделированных систем представлена на рис. 4, *б*. Предложенный метод демонстрирует как минимум не меньшую точность, чем классический, а для сложных моделей более высокую, однако требует существенно меньше времени и количества вычислений функций оценки качества моделей кинетики затухания флуоресценции.

Дополнительно исследуется влияние количества медоидов на точность оценки параметров. Логично предположить, что, как и в методе k-ближайших соседей (в смысле выбора оптимального числа k, ближайших соседей к заданному элементу) [12], выбор одного или нескольких медоидов может быть искажен случайными воздействиями, обусловленными наличием высокого экспериментального шума, что приводит к получению смещенных оценок искомых параметров. В свою очередь рассмотрение слишком большого числа медоидов приводит к грубому усреднению результатов и снижению эффективности метода. Для простых случаев, как в примерах системы 1, увеличение числа медоидов не позволяет улучшить результаты — ошибка ε незначительно изменяется на ~1 % при рассмотрении

одного медоида и среднего значения ε для 5 и 10 медоидов. Для более сложных систем 2 и 3 ε уменьшается более чем на 6 и 8 %. Таким образом, в ходе анализа сложных систем рекомендуется использовать несколько вариантов числа медоидов, оптимальный из которых может быть определен с помощью критерия качества оценки параметров [9].

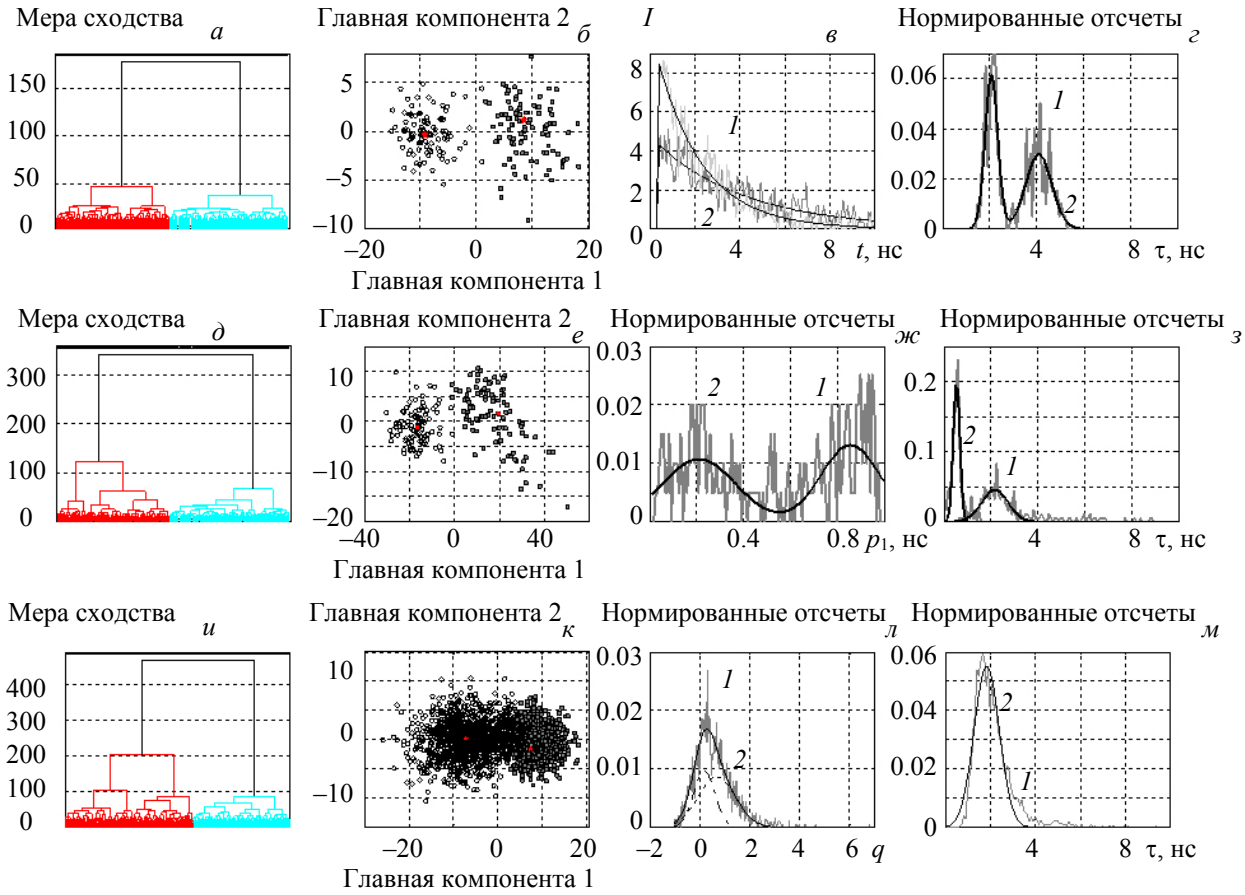

Рис. 3. Результаты анализа смоделированных кривых затуханий флуоресценции систем 1—3 с использованием классического и разработанного методов: *а—г* — система 1, параметры моделирования: $\tau_1 = 2$ и $\tau_2 = 4$ нс, количество кривых 200, $\sigma = 0.1$; *д—з* — система 2, параметры моделирования: $\tau_1 = 0.5$ и $\tau_2 = 2$ нс, их вклады $p_1 = 0.2$, $p_2 = 0.8$ и $p_1 = 0.8$, $p_2 = 0.2$ для двух наборов кривых затуханий, количество кривых 200, $\sigma = 0.1$; *и—м* — система 3, параметры моделирования: $\tau_D = 2$, концентрация акцепторов $q = 1$ и 0.2 для двух наборов кривых затуханий, количество кривых 2000, $\sigma = 0.1$; *а*, *д*, *и* — дендрограммы кластеров кривых затухания флуоресценции; *б*, *е*, *к* — кривые затухания флуоресценции в пространстве первых двух главных компонент, определенные иерархическим кластерным анализом (○, ■), медоиды кластеров (*); размерность осей главных компонент — линейно преобразованное число фотоотсчетов в координатах компонент 1 и 2; *в* — кривые затухания медоидов (*1*) и их аппроксимирующие функции (*2*) в исходном пространстве; *ж*, *л* и *г*, *з*, *м* — результаты аппроксимации гистограмм времен затуханий (*1*), полученных классическим способом, с использованием гауссовых распределений (*2*); оцененные параметры кривых затухания флуоресценции см. в табл. 1—3

Заключение. Предложен метод обработки больших наборов данных кинетических кривых затухания флуоресценции флуорофоров, основанный на использовании алгоритмов интеллектуального анализа данных. Разработанный метод в сравнении с известными позволяет быстрее и точнее определить параметры биофизических и оптических процессов в молекулярных соединениях. Эффектив-

ность алгоритмов предложенного метода проверена в ходе анализа данных, представляющих три системы флуорофоров при различных параметрах проведения вычислительного эксперимента.

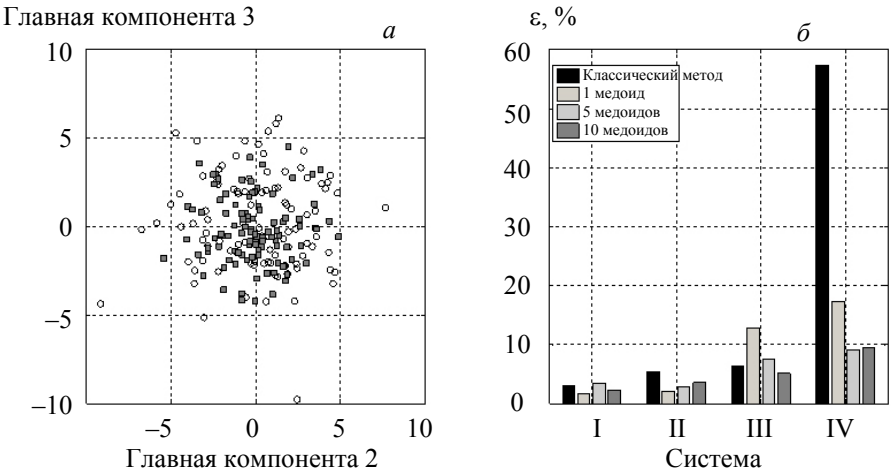

Рис. 4. Кривые затухания флуоресценции системы 1 в пространстве второй и третьей главных компонент (*а*, параметры моделирования указаны в подписи к рис. 3, *а*) и ошибка ε оценки точности восстановления параметров смоделированных кривых затухания флуоресценции систем 1—3 с использованием классического и разработанного методов на основе 1, 5 и 10 медоидов (*б*); I — система 1, $\tau_1 = 2$ и $\tau_2 = 4$ нс, количество кривых 200, $\sigma = 0.1$; II — система 1, $\tau_1 = 1.4$ и $\tau_2 = 2$ нс, количество кривых 200, $\sigma = 0.1$; III — система 2, $\tau_1 = 0.5$ и $\tau_2 = 2$ нс, их вклады $p_1 = 0.2$, $p_2 = 0.8$ и $p_1 = 0.8$, $p_2 = 0.2$ для двух наборов кривых затуханий, количество кривых 200, $\sigma = 0.1$; IV — система 3, $\tau_D = 2$, концентрация акцепторов $q = 1$ и 0.2 для двух наборов кривых затуханий, количество кривых 200, $\sigma = 0.1$

Разработанный метод имеет следующие преимущества над классическим методом анализа данных: позволяет получить более точные оценки времен затухания флуоресценции флуорофоров и демонстрирует надежную сходимость алгоритма оптимизации в область глобального минимума целевой функции; обеспечивает высокую производительность при вычислении целевой функции оценки качества моделей кинетики затухания флуоресценции, что критически важно при использовании имитационных моделей в ходе анализа сложных биомолекулярных систем; предоставляет возможность наглядной визуализации данных в пространстве первых двух/трех главных компонент. Предложенный метод может использоваться в качестве алгоритма поиска начальных приближений, а также для отбора наиболее информативных кривых для традиционных методов анализа. Он может быть применим для анализа больших данных и сложных систем, характеризующихся сверхбольшим набором кривых затухания флуоресценции, анализируемых в условиях ограниченных вычислительных возможностей, обусловленных экономией вычислительных и финансовых ресурсов.

[1] **J. R. Lakowicz.** Principles of Fluorescence Spectroscopy, 3rd ed., Springer, New York (2006)
[2] **A. H. Clayton.** J. Biosci., **43**, N 3 (2018) 463—469
[3] **S. Shashkova, M. C. Leake.** Biosci. Rep., **37**, N 4 (2017); doi: 10.1042/BSR20170031
[4] **D. Phillips.** Proc. Math. Phys. Eng. Sci., **472**, N 2190 (2016) 1—20
[5] **J. P. Angelo, S.-J. Chen, M. Ochoa, U. Sunar, S. Gioux, X. Intes.** J. Biomed. Opt., **24**, N 7, 071602 (2018) 1—20
[6] **E. Wientjes, J. Philippi, J. W. Borst, H. van Amerongen.** Biochim. Biophys. Acta Bioenerg., **1858**, N 3 (2017) 259—265
[7] **A. Boreham, R. Brodwolf, K. Walker, R. Haag, U. Alexiev.** Molecules, **22**, N 1 (2017) E17 (1—18)
[8] Fluorescence Spectroscopy and Microscopy: Methods and Protocols. Methods in Molecular Biology, Eds. Y. Engelborghs, A. J. W. G. Visser, **1076**, Springer Science+Business Media, LLC (2014)
[9] **M. M. Yatskou.** Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Sys-

tems, Wageningen (2001)
[10] **Н. Н. Яцков, В. В. Апанасович, Р. Б. М. Кухорст, А. ван Хук, Т. Й. Схафсма.** Журн. прикл. спектр., **70**, № 3 (2003) 335—339 [**M. M. Yatskou, V. V. Apanasovich, R. B. M. Koehorst, A. van Hoek, T. J. Schaafsma.** J. Appl. Spectr., **70** (2003) 372—377]
[11] **И. В. Станишевский, С. М. Арабей.** Материалы науч.-тех. конф. “Квантовая электроника”, 18—22 ноября 2019 г., Минск, РИВШ (2019) 64—66
[12] **Н. Н. Яцков.** Интеллектуальный анализ данных: пособие, Минск, БГУ (2014)
[13] **Н. Н. Яцков, В. В. Скакун, В. В. Апанасович.** Прикладные проблемы оптики, информатики, радиофизики и физики конденсированного состояния, Минск, НИУ “Ин-т прикл. физ. проблем им. А. Н. Севченко” БГУ (2019) 125—127
[14] **M. Bramer.** Principles of Data Mining, 2nd ed., Springer, London (2013)
[15] **C. C. Aggarwal.** Data Mining: The Textbook, Springer, eBook (2015)
[16] **Н. Н. Яцков, В. В. Апанасович.** Материалы науч.-тех. конф. “Квантовая электроника”, 18—22 ноября 2019 г., Минск, РИВШ (2019) 282—283
[17] **И. Д. Мандель.** Кластерный анализ, Москва, Финансы и статистика (1988)
[18] **М. Б. Лагутин.** Наглядная математическая статистика: уч. пособие, Москва, БИНОМ, Лаборатория знаний (2007)
[19] **I. T. Jolliffie.** Principal Component Analysis, 2nd ed., Springer, New York (2002)
[20] **A. Hyvaerinen, J. Karhunen, O. Erkki.** Independent Component Analysis, New York, John Wiley&Sons Inc. (2001)
[21] **Y. Saeys, I. Inza, P. Larranaga.** Bioinformatics, **23** (2007) 2507—2517
[22] **А. В. Волков, Н. Н. Яцков, В. В. Гринев.** Вестн. БГУ. Математика. Информатика, № 1 (2019) 77—89
[23] **V. Shapaval, J. Brandenburg, J. Blomqvist, V. Tafintseva, V. Passoth, M. Sandgren, A. Kohler.** Biotechnol. Biofuels, **12** (2019) 140 (1—12)
[24] **C. Colabella, L. Corte, L. Roscini, V. Shapaval, A. Kohler, V. Tafintseva, C. Tascini, G. Cardinali.** PLoS One, **12**, N 12 (2017) e0188104 (1—20)
[25] **В. В. Апанасович, О. М. Тихоненко.** Цифровое моделирование стохастических систем, Минск, Университетское (1986)
[26] **T. A. Roelofs, C. H. Lee, A. R. Holzwarth.** Biophys. J, **61**, N 5 (1992) 1147—1163
[27] **A. V. Digris, E. G. Novikov, V. V. Skakun, V. V. Apanasovich.** Method. Mol. Biol., **1076** (2014) 257—277
[28] **В. В. Апанасович, Е. Г. Новиков, Н. Н. Яцков, Р. Б. М. Кухорст, Т. Й. Схафсма, А. ван Хук.** Журн. прикл. спектр., **66**, № 4 (1999) 549—552 [**V. V. Apanasovich, E. G. Novikov, N. N. Yatskov, R. B. M. Koehorst, T. J. Schaafsma, A. van Hoek.** J. Appl. Spectr., **66** (1999) 613—616]
[29] **В. В. Апанасович, Е. Г. Новиков, Н. Н. Яцков.** Журн. прикл. спектр., **67**, № 5 (2000) 612—618 [**V. V. Apanasovich, E. G. Novikov, N. N. Yatskov.** J. Appl. Spectr., **67** (2000) 842—851]
[30] **Н. Н. Яцков, Е. В. Лисица.** Интеллектуальный анализ данных: методические указания к лабораторным работам, Минск, БГУ (2019)