ВЫЧИСЛИТЕЛЬНАЯ ПЛАТФОРМА FLUORSIMSTUDIO ДЛЯ ОБРАБОТКИ КИНЕТИЧЕСКИХ КРИВЫХ ЗАТУХАНИЯ ФЛУОРЕСЦЕНЦИИ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Н. Н. Яцков*, В. В. Апанасович

УДК 535.37

Белорусский государственный университет,
220030, Минск, Беларусь; e-mail: yatskou@bsu.by

(Поступила 11 февраля 2021)

Создана вычислительная платформа FluorSimStudio для обработки кривых затухания флуоресценции в молекулярных системах, реализующая концепцию комплексного подхода к анализу экспериментальной информации на основе методов имитационного моделирования и интеллектуального анализа данных. Комплексный анализ включает в себя разделение кривых затухания флуоресценции на кластеры по степени близости в некоторой мере сходства, нахождение медианных представителей кластеров (медоидов), применение метода снижения размерности данных и отображение экспериментальных данных в двухмерном пространстве. Анализ кривых затуханий медоидов осуществляется с использованием аналитических или имитационных моделей оптических процессов, протекающих в молекулярных системах. Визуализация кластеров данных в исходном и преобразованном временном пространствах проводится с целью обеспечения интерактивного взаимодействия. Предложена схема функционирования платформы, обоснован выбор программных средств для обеспечения высокой производительности вычислений, реализовано веб-приложение платформы (https://dsa-cm.shinyapps.io/FluorSimStudio), приведены результаты сравнительного анализа алгоритмов имитационного моделирования. Работоспособность вычислительной платформы подтверждена примерами анализа наборов данных, представляющими системы свободных флуорофоров и при наличии процесса переноса энергии электронного возбуждения по Фёрстеру. Вычислительная платформа является открытой системой и допускает постоянное добавление моделей комплексного анализа с учетом разработки новых алгоритмов имитационного моделирования процессов переноса энергии в молекулярных системах, регистрируемых с помощью систем флуоресцентной спектроскопии с временным разрешением.

Ключевые слова: *флуоресцентная спектроскопия, затухание флуоресценции, имитационное моделирование, интеллектуальный анализ данных, вычислительная платформа, разработка программных средств, программирование на R.*

Herein, a computational platform FluorSimStudio was developed for processing fluorescence decay curves in molecular systems, which implements the concept of complex analysis of experimental information based on the simulation modelling and data mining methods. Data analysis includes partitioning the fluorescence decay curves into clusters according to the degree of likeness to some measure of similarity, finding the median cluster members (medoids), applying the data reduction method and visualizing the experimental data in a two-dimensional space. Analysis of the decay curves is carried out by the analytical or simulation models of optical processes occurring in molecular systems. The visualization of data clusters in the original and transformed time spaces is done with the aim of user interaction. A functional scheme of the

---

COMPUTATIONAL PLATFORM FLUORSIMSTUDIO FOR PROCESSING THE KINETIC CURVES OF FLUORESCENCE DECAY USING SIMULATION MODELLING AND DATA MINING ALGORITHMS

M. M. Yatskou*, V. V. Apanasovich (*Belarusian State University, Minsk, 220030, Belarus; e-mail: yatskou@bsu.by*)

platform is proposed, the choice of software for ensuring high computing performance is substantiated, a web application of the platform is implemented (https://dsa-cm.shinyapps.io/FluorSimStudio), and the results of a comparative analysis of the simulation algorithms are presented. The performance of the computational platform was confirmed by examples of the analysis of data sets representing systems of free fluorophores and in the presence of the Förster electronic excitation energy transfer process. The computational platform is an open system and allows permanent addition of complex analysis models, taking into account the development of new algorithms for modelling the energy transfer processes in molecular systems, studied with the use of time-resolved fluorescence spectroscopy systems.

Keywords: *fluorescence spectroscopy, fluorescence decay, simulation modelling, data mining, computational platform, software development, R programming.*

Введение. Современные методы флуоресцентной спектроскопии с временным разрешением позволяют регистрировать большие наборы кинетических кривых затухания флуоресценции биофизических систем [1, 2]. Первоочередной задачей является разработка эффективных моделей, методов и программных средств обработки серий данных флуоресцентной спектроскопии. Предложены различные алгоритмы анализа и моделирования данных [3—6], среди которых можно выделить метод обработки больших наборов кинетических кривых затухания флуоресценции молекул с использованием алгоритмов имитационного моделирования и интеллектуального анализа данных. Его применение позволяет повысить точность оцененных параметров биофизических и оптических процессов, протекающих в исследуемых молекулярных системах [6]. Разработаны специализированные и общего назначения программные средства и продукты, как коммерческие, так и находящихся в свободном доступе, предназначенные для статистической обработки, анализа и имитационного моделирования кривых затухания флуоресценции. Однако единые комплексные программные средства обработки больших наборов данных с использованием методов имитационного моделирования и интеллектуального анализа отсутствуют. Разработка программной платформы имитационного моделирования и интеллектуального анализа кривых затухания флуоресценции в молекулярных системах, используемой при применении комплексного подхода в ходе анализа различных биофизических систем в экспериментальных исследованиях, является важной и актуальной задачей.

В настоящей работе предложена цифровая платформа программных средств для обеспечения комплексного подхода к анализу и моделированию оптических процессов в молекулярных системах, регистрируемых с помощью экспериментальных методов флуоресцентной спектроскопии с временным разрешением. Вычислительная платформа организована по примеру открытых проектов сетевых ресурсов CRAN (https://cran.r-project.org), R-Forge (https://r-forge.r-project.org), Bioconductor (https://www.bioconductor.org), Github (https://github.com). Она является средой программирования и имитационного моделирования, содержит обновляемые и дополняемые библиотеки аналитических и имитационных моделей оптических процессов в молекулярных системах, встроенные инструменты методов интеллектуального анализа данных и оценки качества анализа и моделирования, предоставляет научному сообществу возможности для разработки новых алгоритмов и имитационных моделей.

Обоснование выбора программной платформы. Под вычислительной платформой понимаем интеллектуальный программный ресурс или среду программирования, предназначенную для решения задач моделирования и анализа больших экспериментальных данных биофизических исследований в прикладной спектроскопии. Платформа включает в себя среду программирования, интегрированные языки программирования, программные средства автоматизации, отладки кодов и создания интерфейса приложения, модели объектов исследования, методы анализа и визуализации данных, оценки качества анализа и достоверности моделей. Выбор оптимальной программной платформы в первую очередь подразумевает выбор среды программирования и средств разработки интерфейса для взаимодействия с пользователем.

Для реализации программного обеспечения используются различные вычислительные платформы и технологии программирования. В большинстве публикаций по сравнительному анализу пакетов открытого доступа в области интеллектуального анализа данных нет явного лидера. В настоящее время активно используется большое количество программных средств, среди которых можно выделить WEKA, Tanagra, Rapid Miner, KNIME, Orange, Java-, Python- и R-проекты, а также платформы, реализованные с помощью высокопроизводительных языков программирования C++ и Scala [7—13]. Достоинствами того или иного программного ресурса являются вычислительная производительность, широкий набор подключаемых библиотек статистического анализа, кроссплатформенность, возмож-

ности разработки пользовательских интерфейсов, выполнения распараллеленных вычислений, работы напрямую с существующими базами и хранилищами данных. К основным недостаткам следует отнести отсутствие универсальности, существенные требования к вычислительным ресурсам, ограничение интегрированности вышеуказанных свойств в едином формате. Наиболее перспективные проекты организации цифровой среды — Scala-, Python- и R-платформы. Платформа на основе языка Scala (например, Apache Hadoop [14]) предназначена для анализа больших данных в производственных проектах и используется при решении задач промышленного программирования. Python-приложения нацелены на решение общих задач инженерии и анализа данных с акцентом на нейросетевые подходы и программирование. R-проекты разрабатывается в первую очередь с целью оптимизации и достоверности прикладного статистического анализа, включающего в себя подходы с использованием методов классического и интеллектуального анализа данных. Остановимся подробнее на рассмотрении среды R.

Основные преимущества среды статистического программирования R — наличие оптимизированных структур представления объектов данных, значительно упрощающее обработку данных, оптимизация инструментов программирования и реализации алгоритмов вычислений (в смысле минимизации внесения ошибок в программный код), возможность использования огромного набора алгоритмов обработки, статистического и интеллектуального анализа данных, разнообразных вычислительных ресурсов научного сообщества [15, 16]. Главный недостаток — невысокая вычислительная производительность в базовом варианте размещения среды, что особенно критично при работе с большими наборами данных и разработке имитационных моделей. Указанное ограничение можно частично или полностью устранить с помощью подключения программных кодов высокопроизводительных языков программирования Scala, Java, C++ (пакеты rscala, rjava, Rcpp, inline), процедур распараллеливания вычислений (управляются пакетами parallel, Rmpi, snow, snowfall), дополнительных пакетов эффективной обработки больших данных (readr, LaF, data.table, ff, bigmemory) и использования сторонних программных ресурсов (библиотек Microsoft R Open и Intel Math Kernel Library, платформы анализа больших данных H2O [17], систем Apache Hadoop и Spark [18], с помощью пакетов h20, Rhadoop и SparkR).

Важный вопрос — разработка интерфейса программного приложения. Наиболее популярными пакетами для разработки пользовательских интерфейсов программных приложений, интегрирующими R-коды, являются gWidgets, rpanel, svDialogs, RGtk2, qtbase, tcltk. Новое направление в разработке R-приложений для анализа биофизических систем [19—21] связано с созданием “реактивных” веб-интерфейсов с использованием пакета Shiny и последующим размещением программной реализации на ресурсе shinyapps.io, предоставляемом разработчиками открытого программного обеспечения RStudio [22]. Достоинство данного подхода — возможность удаленной работы с веб-приложением широкой научной аудитории пользователей в режиме онлайн через глобальную сеть Internet. Для реализации программного приложения в работе выбраны вычислительная среда R и пакет Shiny для создания веб-интерфейса разработанного приложения.

Разработка схемы функционирования платформы. Программная платформа интегрирует схему исследования некоторого биофизического процесса или молекулярного соединения с использованием комплексного подхода на основе применения методов имитационного моделирования и интеллектуального анализа данных [6]. Решается задача анализа набора кривых затухания флуоресценции молекулярных систем с целью определения параметров оптико-физических процессов, протекающих в исследуемых системах. Схема методики интегрированного в платформу анализа данных флуоресцентной спектроскопии с временным разрешением представлена на рис. 1. Рассмотрим основные этапы анализа данных.

Платформа предназначена для анализа экспериментальных или смоделированных данных. Загрузка и графическое представление данных осуществляется в блоке 1. Визуальная оценка двухмерных и трехмерных кривых затухания флуоресценции позволяет предопределить выбор той или иной математической модели описания кинетики затухания, сделать предположение о количестве кластеров данных, на основе зашумленности данных ограничить выбор мер вычисления сходства кривых затуханий. Осуществляется выбор обрабатываемых данных, предлагается анализ экспериментальных или смоделированных наборов.

Моделирование и визуализация кривых затухания флуоресценции проводятся в блоке 2. Рассматриваются интегрированные модели оптических процессов. Имитационное моделирование осуществляется с использованием алгоритмов Монте-Карло [23]. Вводимыми характеристиками моде-

лирования являются вид и параметры модели, количество кривых и число имитаций, тип и параметры инструментальной функции отклика прибора. Двухмерная и трехмерная визуализации кривых затухания предназначены для экспертного анализа смоделированных кривых, исследования поведения моделей при изменении их параметров, ручного подбора наиболее оптимальных параметров моделирования, таких как число имитаций и кривых затухания, а также начальных приближений параметров для последующего точного определения в ходе интеллектуального анализа данных с использованием математических моделей. Новые и улучшенные модели оптико-физических процессов в молекулярных системах могут быть разработаны и интегрированы в программную среду.

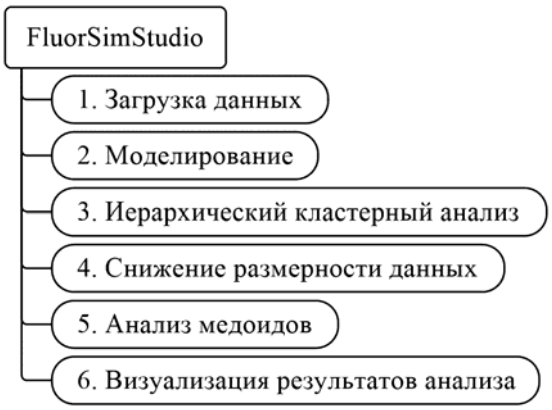

Рис. 1. Схема основных этапов исследования флуоресценции молекулярных соединений с использованием алгоритмов математического моделирования и интеллектуального анализа данных

В блоке 3 выполняется кластерный анализ кривых затухания флуоресценции в пространстве исходных признаков, представленных числом фотоотчетов во временных каналах гистограмм регистрируемых фотонов. Определяются кластеры кривых затуханий флуоресценции в некоторой мере сходства (евклидово, Минковского, манхэттенское, максимальное (или Чебышева, максимум модуля разности компонент векторов) или Канберра расстояния). Число кластеров определяется интуитивно, автоматически по дендрограмме кривых затухания флуоресценции, построенной на основе меры связывания кластеров (Уорда, ближайшего соседа, дальнего соседа или средней связи), или на основе статистического критерия [24, 25]. Вычисляются медианные представители кластеров — медоиды, кривые затухания, имеющие наименьшие средние расстояния до остальных объектов соответственных кластеров.

Снижение размерности данных проводится в блоке 4. Учет большой группы малоинформативных признаков, представленных количеством фотоотчетов в заданные моменты времени, приводит к затруднению анализа данных, а именно к 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, перспективно 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, среди которых наиболее широко известен метод главных компонент [26]. Выполняется преобразование кривых затуханий флуоресценции с помощью метода главных компонент. Задается доля относительной вариации, приходящаяся на главные компоненты, ограничивающая число компонент. Отбираются главные компоненты, соответствующие заданной вариации в данных (например, 0.95). Строится диаграмма долей вариации первых десяти главных компонент, по которой оценивается вклад в общую дисперсию в данных. Отображаются кластеры и их медоиды на диаграмме рассеяния первых двух главных компонент. Медоиды вычисляются в пространстве исходных признаков либо в пространстве главных компонент, объясняющих заданную долю изменчивости. Если кластеры данных не разделяются, то можно полагать, что присутствует только один вид молекулярных соединений. Иначе допускается наличие нескольких форм молекулярных

соединений (флуорофоров). Для удобства визуального контроля разделимости кластеров строятся гистограммы частот на оси первых трех главных компонент. Для хорошей разделимости кластеров характерно наличие многомодального вида распределений гистограмм.

В блоке 5 выполняется анализ медоидов кластеров для точного определения параметров молекулярных соединений с использованием алгоритма оптимизации и математических моделей. Для аппроксимации кинетических кривых затухания флуоресценции, представленных найденными медоидами, применяются аналитические и имитационные модели описания фотофизических процессов [6, 23]. Для оптимального подбора параметров математических моделей в ходе аппроксимации экспериментальных данных используются методы оптимизации. В настоящей работе выбран метод Нелдера—Мида [27], который не учитывает взятия производной целевой функции, что существенно упрощает применение имитационных моделей в процедуре оценки параметров. Наилучшее приближение выбирается по критерию (или набору критериев), определяющему степень отклонения теоретической модели от экспериментальных данных. Как правило, такой критерий представляется аналитически в виде функции экспериментальных и теоретических данных, вид которой определяется областью применения, непосредственным методом моделирования и условиями проведения эксперимента. В наших экспериментах рассматриваются нормированный критерий χ^2, диаграммы взвешенных остатков и их автокорреляционной функции [23]. χ^2-Подобный критерий:

$$\chi^2(a) = \frac{1}{\nu}\sum_{i=1}^{n}\frac{[E(t_i) - I(t_i,a)]^2}{w(t_i)}, \tag{1}$$

$$w(t_i) = \mathrm{var}[E(t_i) - I(t_i,a)] = \mathrm{var}[E(t_i)] + \mathrm{var}[I(t_i,a)], \tag{2}$$

где n — число каналов многоканального анализатора; $w(t_i)$ — весовой фактор; $E(t_i)$ — экспериментальная кривая затухания флуоресценции, представленная числом фотоотчетов в каналах; $I(t_i,a)$ — теоретическая гистограмма, представленная числом фотоотчетов N смоделированных фотонов в каналах и характеризующаяся наименьшим числом p параметров $a = \{a_1, a_2, \ldots, a_p\}$, описывающих моделируемую систему; $\nu = n - p - 1$ — число степеней свободы, var — дисперсия. Измеряемая интенсивность затухания флуоресценции является сверткой функции отклика образца и функции вспышки лазера, обычно представляемой конечной функцией отклика аппаратуры. Математически свертка для интенсивности затухания $I(t,a)$ 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 отклика аппаратуры, моделируется в виде прямоугольного импульса; $i(t,a)$ — функция отклика образца, определяется математической моделью исследуемой системы. Предполагается, что $E(t_i)$ и $I(t_i,a)$ статистически не зависимы, а модель $I(t_i,a)$ является наилучшей аппроксимацией экспериментальных данных $E(t_i)$. В соответствии с выбранным критерием, наилучшим приближением, определяемым набором оцененных параметров $\hat{a}$, является то, которое обеспечивает минимум критерия. Для приведенных условий выражение критерия χ^2 может быть непосредственно использовано в качестве целевой функции в процедуре оценки параметров моделей.

Для углубленной оценки качества проведенного анализа используют дополнительные критерии, к которым относятся диаграммы взвешенных остатков и их автокорреляционной функции, позволяющие визуально оценивать степень совпадения измеряемых данных и теоретической модели. Взвешенные остатки можно рассчитать по формуле

$$R(t_i) = \frac{E(t_i) - I(t_i,a)}{\sqrt{w(t_i)}}, \quad i = 1,2,\ldots,n. \tag{4}$$

Если экспериментальная и теоретическая кривые затухания согласуются, то взвешенные остатки распределены нормально около нулевого значения. Если между экспериментальной и теоретической кривыми затухания флуоресценции имеются незначительные систематические отклонения, не различимые по графику взвешенных остатков, то для их нахождения применяется автокорреляционная функция взвешенных остатков:

$$ACF(t_k) = \frac{\dfrac{1}{n-k+1}\displaystyle\sum_{i=1}^{n-k+1} R(t_i)R(t_{i+k-1})}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} R^2(t_i)}, \quad k = 1,2,\ldots,n/2. \tag{5}$$

Если отклонения между экспериментальными и теоретическими кривыми затухания случайны, что характерно для наилучшего выбора теоретической модели и ее параметров, автокорреляционная функция взвешенных остатков осциллирует с малой амплитудой около нуля.

Визуализация результатов и анализ графических образов оцененных кластеров кривых затухания проводятся с целью интерпретации, объяснения, улучшения понимания объекта исследования и его поведения (блок 6). Кластеры кривых затухания строятся в пространстве трех главных компонент, исходном (временном) пространстве и координатах главных компонент, объясняющих заданную долю вариации в данных. Представление диаграммы трех главных компонент, интерактивной для взаимодействия с пользователем, позволяет визуально оценить близость найденных кластеров и их формы, расположение отдельных кривых, влияние экспериментальных эффектов. Диаграммы набора информативных компонент позволяют определить кластеры данных для возможной оценки параметров моделей в пространстве главных компонент. Последнее помогает повысить точность оценки параметров за счет снижения шума в кривых затухания вследствие исключения неинформативных компонент, описывающих экспериментальный шум. Процедура оценки параметров моделей в пространстве главных компонент может быть дополнительно реализована в платформе. Интерактивная диаграмма кластеров кривых затуханий во временном пространстве позволяет качественно исследовать группы кривых затуханий.

Программная реализация платформы FluorSimStudio. Для практической реализации комплексного подхода выбрана вычислительная платформа на базе ресурсов языков программирования R и C++. Прототипом для макета платформы является программный пакет RNAexploreR [7]. Структура платформы представлена на рис. 2.

Запуск проекта осуществляется на сервере R, размещенном на сетевом ресурсе, например shinyapps.io. Для реализации имитационных моделей предлагается использовать язык программирования C++. Выбор и разработка алгоритмов анализа данных осуществляются путем непосредственного программирования или подключения готовых пакетов интеллектуального анализа данных (ИАД), предоставляемых научным сообществом разработчиков через открытые проекты CRAN, Bioconductor, Github. Работа пользователя осуществляется через веб-приложение. В структуре вычислительного подхода платформа интегрирует реализации имитационных моделей, алгоритмов анализа и оценки качества анализа, предоставляет вычислительные средства для применения разработанных имитационных моделей и методов к анализу наборов данных, инструменты для оценки его качества, визуализации и интерпретации данных.

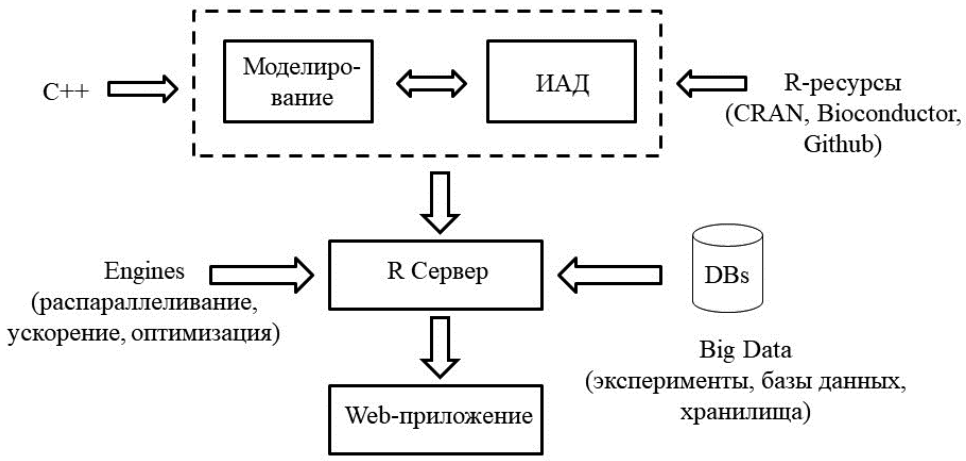

Рис. 2. Вычислительная платформа комплексного анализа на базе ресурсов языка программирования R

Программная реализация платформы FluorSimStudio организована с помощью R пакета Shiny и содержит набор функций, интегрирующих методику комплексного подхода к анализу данных. Веб-приложение размещено на ресурсе https://dsa-cm.shinyapps.io/FluorSimStudio. Пример окон интерфей-

са пакета представлен на рис. 3. Главное окно интерфейса состоит из девяти панелей, соответствующих шести этапам анализа: загрузки и моделирования данных, снижения размерности данных с помощью метода главных компонент, анализа медоидов, визуализации и интерпретации результатов, информации об авторах разработки, инструкции по использованию вычислительного ресурса.

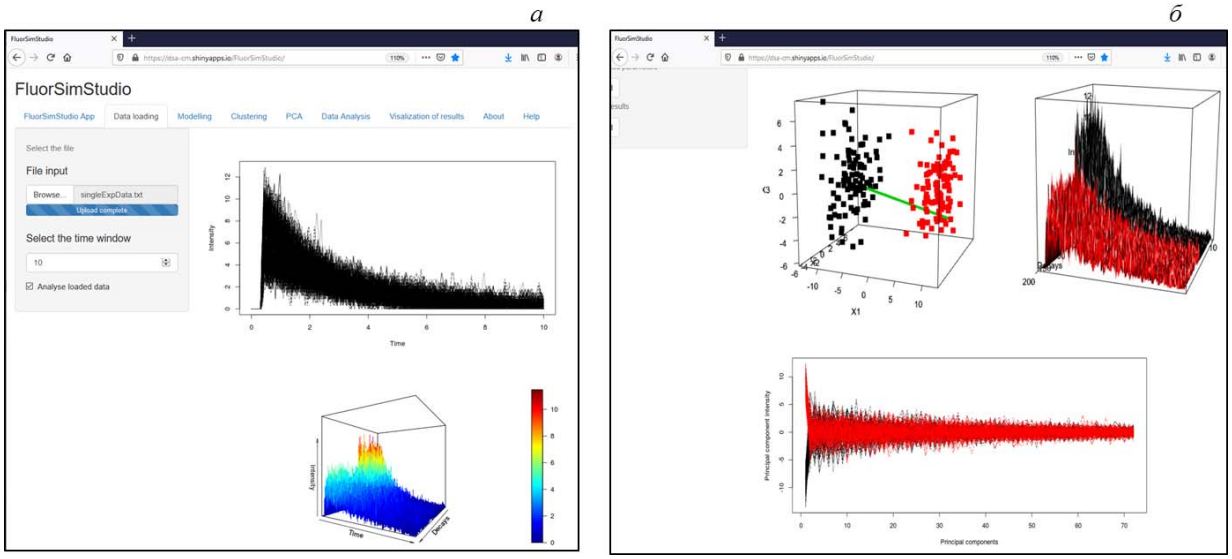

Рис. 3. Окна интерфейса веб-приложения FluorSimStudio: пример этапов загрузки данных (*а*) и визуализации результатов анализа (*б*) кривых затухания флуоресценции

Реализованы имитационные модели кинетических кривых затухания флуоресценции двух систем молекул в некотором абстрактном эксперименте (системы 1 и 2): 1) двух люминесцирующих молекулярных мономеров, пространственно разделенных на некоторой поверхности, 2) люминесцирующего молекулярного донора, окруженного нелюминесцирующими молекулами акцептора [6]. Первая система проста для понимания и анализа, позволяет подтвердить применимость разработанных алгоритмов. Вторая система — более сложная в первую очередь для моделирования и анализа данных, поверхности целевой функции которой имеют “оврагоподобный” характер, что существенно затрудняет поиск глобального минимума, соответствующего наилучшим оценкам параметров моделей. Смоделированные данные позволяют качественно и количественно оценить работоспособность платформы. Рассмотрим математические модели исследуемых систем.

Модель системы 1. Данная система описывается одноэкспоненциальной моделью затухания. Экспоненциальная модель — наиболее простая модель интенсивности затухания флуоресценции, используется для математического описания растворов низких концентраций невзаимодействующих молекул [28], интенсивность которых

$$i(t,i_0,\tau) = i_0 \exp\{-t/\tau\}, \tag{6}$$

где τ и i_0 — время затухания и интенсивность флуоресценции в момент времени $t = 0$ (относительно импульса возбуждения). Алгоритм имитационного моделирования приведен в [6].

Модель системы 2. Модель стрэтч, или “растянутой” экспоненты, представляет собой аналитическое описание кинетики затухания интенсивности флуоресценции молекул доноров в донорно-акцепторной системе в присутствии переноса энергии электронного возбуждения по Фёрстеру [28]. Флуоресценция донора в трехмерном пространстве может быть описана выражением

$$i(t,i_0,q,\tau_D) = i_0 \exp\{-t/\tau_D - q(t/\tau_D)^{1/2}\}, \tag{7}$$

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 концентрации акцепторов; τ_D — время затухания флуоресценции молекул доноров. Имитационная модель реализована на основе метода Неймана [6].

Вычислительные процедуры запрограммированы с использованием R-функций dist, hclust, cluster, eigen, optim, интегрирующих алгоритмы иерархического кластерного анализа, методов главных компонент и Нелдера—Мида. Начальные приближения параметров моделей задаются в окне

ввода параметров. Следует отметить, что пользователь, работая в среде R, может создавать собственные модели, импортировать из существующих проектов или адаптировать представленные инструменты пакета для собственных задач.

Тестирование программных средств проведено на ПК с характеристиками DualCore Intel Pentium E5700, 3000 MHz, 8156 MB DDR3-1333 RAM.

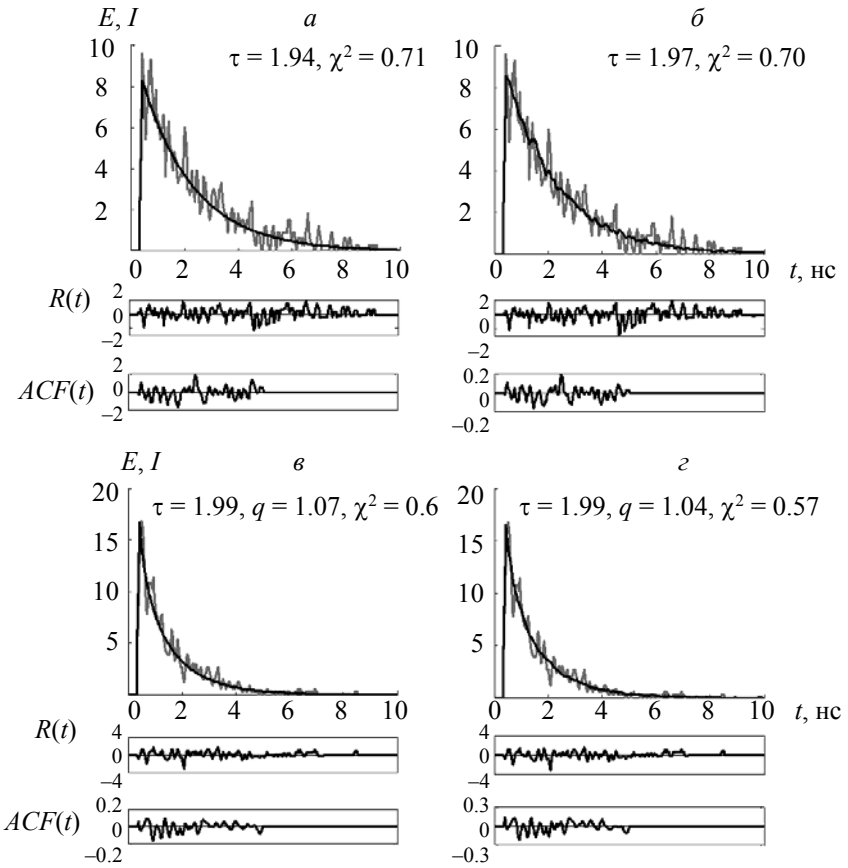

Рис. 4. Результаты анализа медоидов кластеров условно-экспериментальных кривых затухания флуоресценции молекулярных систем 1 и 2 с использованием аналитических (*а*, *в*) и имитационных (*б*, *г*) моделей; приведены оцененные параметры моделей при минимуме критерия χ^2; экспериментальные кривые затухания флуоресценции: *а* и *б* — система 1, параметры моделирования: времена жизни $\tau_1 = 2$ нс и $\tau_2 = 4$ нс для двух наборов кривых затуханий; *в* и *г* — система 2, время жизни доноров $\tau_D = 2$ нс, концентрация акцепторов $q = 1$ и 0.2 для двух наборов кривых затуханий; длина интервала наблюдения 10 нс, число временных каналов гистограмм 256, количество кривых 200, число смоделированных фотонов в экспериментальных кривых 500, в теоретических $5 \cdot 10^4$, функция отклика прибора моделируется в виде прямоугольного импульса $\sim 10^{-2}$ интервала наблюдения

Комплексный подход требует использования затратных программных ресурсов для процедур имитационного моделирования. Язык C++, интегрированный в пакет Rcpp, применяется для ускорения имитационного моделирования. Важной частью имитационной модели является генерация псевдослучайных чисел, базовый элемент которой — генератор равномерной случайной величины. Программные реализации генератора равномерной случайной величины интегрированы в функциях runif (пакет stats среды R), runif (пакет Rcpp) и rand (C++ библиотека cstdlib). Обоснованный выбор наиболее оптимальной функции программной реализации датчика псевдослучайных чисел поможет снизить время имитационного моделирования. Для проверки эффективности алгоритмов моделирования, реализованных в среде R и на языке C++, проведен вычислительный эксперимент по исследованию времени моделирования одно- и стрэтч-экспоненциальной моделей затухания флуоресценции. Дополнительно, в качестве эталонной системы рассмотрено имитационное моделирование в среде

Matlab на базе функции rand генерации реализаций равномерной случайной величины. Регистрировалось время моделирования для числа фотонов 10^8. Выполнено 10 циклов моделирования, по результатам которых вычислено среднее время моделирования. Параметры моделирования: $\tau = 2$ нс, $\tau_D = 2$ нс, $q = 1$, число каналов 256, длина интервала наблюдения 10 нс.

Генераторы runif среды R (8.21 и 60.80 с) и rand C++ (8.65 и 60.49 с) самые быстрые и позволяют сократить время моделирования одно- и стрэтч-экспоненциального затухания в сравнении с генератором runif пакета Rcpp (14.61 и 82.03 с). C++ программные реализации алгоритмов моделей быстрее в пять раз или более, чем их Matlab- (285 и 325 с) и R (48 и 1931 с) аналоги. R-реализации алгоритмов имитационного моделирования самые медленные и фактически не пригодны для использования в комплексном подходе. Последнее утверждение в принципе очевидно, однако получено количественное подтверждение. Принято решение использовать датчик случайных чисел функции rand C++ библиотеки cstdlib.

Эффективность работы программного приложения FluorSimStudio проверена на примерах анализа наборов кривых затухания флуоресценции систем люминесцирующих флуорофоров, сгенерированных с использованием моделей молекулярных систем 1 и 2. Полученные результаты хорошо согласуются с опубликованными ранее для аналитических моделей одно- и стрэтч-экспоненциального законов затухания флуоресценции [6]. Комплексный анализ с использованием имитационных моделей и интеллектуального анализа данных позволяет восстановить параметры оптических процессов, установленные при генерации условно-экспериментальных данных. Результаты анализа наборов кривых затухания флуоресценции для молекулярных систем 1 и 2, полученные с использованием FluorSimStudio, представлены на рис. 4.

Заключение. Разработана цифровая платформа FluorSimStudio, которая является программной реализацией комплексного подхода для анализа и моделирования оптических процессов в биофизических системах. Основное программное обеспечение — онлайн-приложение, состоит из нескольких модулей, соответствующих различным этапам моделирования и анализа данных.

Платформа имеет следующие преимущества: реализует концепцию комплексного анализа данных с использованием имитационного моделирования и интеллектуального анализа данных; обеспечивает высокую производительность обработки больших наборов кривых затухания флуоресценции, что критически важно при использовании имитационных моделей в изучении сложных биомолекулярных систем; представляет возможность наглядной визуализации данных в пространстве первых двух-трех главных компонент; расширяется за счет включения новых имитационных моделей и алгоритмов анализа данных; онлайн-вариант размещен на сервере, может использоваться в образовательном процессе и для исследования экспериментальных систем; вычислительная эффективность может быть увеличена за счет подключения программных средств для высокопроизводительных вычислений и анализа больших данных (например, ресурсов H2O, Apache Hadoop, Spark).

Платформу можно использовать в качестве инструмента для поиска начальных приближений параметров моделей и изучения их чувствительности к внешним возмущениям, в том числе в условиях высокого экспериментального шума, исследования пределов разделимости кластеров кривых затухания флуоресценции и последующего планирования отдельных экспериментов, а также для обучения основам анализа данных флуоресцентной спектроскопии с временным разрешением. Она может быть рекомендована для анализа данных сложных систем, характеризующихся сверхбольшим набором кривых затухания флуоресценции, анализируемых в условиях ограниченных возможностей, обусловленных экономией вычислительных и финансовых ресурсов.

[1] **R. R. Choubeh, L. Bar-Eya, Y. Paltiel, N. Keren, P. C. Struik, H. van Amerongen.** Photosynth. Res., **143** (2020) 13—18

[2] **L. Michels, V. Gorelova, Y. Harnvanichvech, J. W. Borst, B. Albada, D. Weijers, J. Sprakel.** Proc. Natl. Acad. Sci. USA, **117**, N 30 (2020) 18110—18118

[3] Fluorescence Spectroscopy and Microscopy: Methods and Protocols. Methods in Molecular Biology, Eds. Y. Engelborghs, A. J. W. G. Visser, Springer Science+Business Media, LLC (2014) 1076

[4] **J. T. Smith, R. Yao, N. Sinsuebphon, A. Rudkouskaya, N. Un, J. Mazurkiewicz, M. Barroso, P. Yan, X. Intes.** Proc. Natl. Acad. Sci. USA, **116**, N 48 (2019) 24019—24030

[5] **W. M. J. Franssen, F. J. Vergeldt, A. N. Bader, H. van Amerongen, C. Terenzi.** J. Phys. Chem. Lett., **11**, N 21 (2020) 9152—9158

[6] **Н. Н. Яцков, В. В. Скакун, В. В. Апанасович.** Журн. прикл. спектр., **87**, № 2 (2020) 322—333

[**M. M. Yatskou, V. V. Skakun, V. V. Apanasovich.** J. Appl. Spectr., **87**, N 2 (2020) 333—344]
[7] **Н. Н. Яцков, В. В. Скакун, В. В. Гринев.** Информатика, **16**, № 4 (2019) 7—24
[8] **J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik, B. Zupan.** J. Machine Learn. Res., **14** (2013) 2349—2353
[9] **M. F. Hornick, E. Marcadé, S. Venkayala.** Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for Architecture, Design, and Implementation. Morgan Kaufmann Publishers Inc., San Francisco (2006)
[10] **F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay.** J. Machine Learn. Res., **12** (2011) 2825—2830
[11] **D. Schmidt, W.-C. Chen, M. A. Matheson, G. Ostrouchov.** Big Data Res., **8** (2016) 1—11
[12] **T. Masters.** Data Mining Algorithms in C++. Data Patterns and Algorithms for Modern Applications, Apress, eBook (2018)
[13] **J. M. Abuín, N. Lopes, L. Ferreira, T. F. Pena, B. Schmidt.** PLoS One, **15**, N 10 (2020) e0239741, doi: 10.1371/journal.pone.0239741.
[14] Apache Software Foundation. Apache Hadoop, http://hadoop.apache.org
[15] R Core Team. R: A Language and Environment for Statistical Computing. Foundation for Statistical Computing, Vienna, Austria (2020), http://www.R-project.org
[16] **R. Gentleman, V. J. Carey, D. M. Bates.** Genome Biology, **5**, N 10 (2004) R80, doi: 10.1186/gb-2004-5-10-r80
[17] H2O.ai. (2020) H2O: Scalable Machine Learning Platform. Version 3.30.0.6. https://github.com/h2oai/h2o-3
[18] **M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica.** Commun. ACM, **59**, N 11 (2016) 56—65
[19] **T. Zhu, H. Chen, X. Yan, Z. Wu, X. Zhou, Q. Xiao, W. Ge, Q. Zhang, C. Xu, L. Xu, G. Ruan, Z. Xue, C. Yuan, G.-B. Chen, T. Guo.** Bioinform. (**2021**) btaa1088, doi: 10.1093/bioinformatics/btaa1088
[20] **V. Yuan, D. Hui, Y. Yin, M. S. Peñaherrera, A. G. Beristain, W. P. Robinson.** BMC Genomic., **22**, N 1 (2021), doi: 10.1186/s12864-020-07186-6
[21] **J. Lu, S. L. Salzberg.** PLoS Comput Biol., 1**6**, N 12 (2020) e1008439, doi: 10.1371/journal.pcbi.1008439
[22] RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston (2020), http://www.rstudio.com
[23] **M. M. Yatskou.** Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Systems, Wageningen (2001)
[24] **Н. Н. Яцков.** Интеллектуальный анализ данных: пособие, Минск, БГУ (2014)
[25] **H. Shimodaira.** Annal. Statist., **32** (2004) 2616—2641
[26] **T. Jolliffie.** Principal Component Analysis, Springer, New York (2002)
[27] **J. A. Nelder, R. Mead.** Comput. J., **8** (1965) 308—313
[28] **J. R. Lakowicz.** Principles of Fluorescence Spectroscopy, Springer, New York (2006)