Оптика и спектроскопия

Optics and spectroscopy

УДК 535.37

РАЗРАБОТКА КОМПЛЕКСНОГО ПОДХОДА, ОСНОВАННОГО НА МЕТОДАХ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ, ДЛЯ ИССЛЕДОВАНИЯ СИСТЕМ ПРИКЛАДНОЙ ФЛУОРЕСЦЕНТНОЙ СПЕКТРОСКОПИИ

Н. Н. ЯЦКОВ[1)]***, В. В. АПАНАСОВИЧ***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Для исследования биомолекулярных соединений в системах обеспечения прикладной флуоресцентной спектроскопии предлагается применять комплексный подход, основанный на методах имитационного моделирования и интеллектуального анализа данных и включающий имитационные модели физических процессов, методы и алгоритмы интеллектуального анализа данных, программные средства автоматизации исследования молекулярных и клеточных систем. Идея комплексного подхода состоит в использовании имитационного моделирования биофизических процессов, протекающих в объекте исследования, отборе наиболее информативных экспериментальных данных, определении характеристик объекта с применением алгоритмов интеллектуального анализа данных. Эффективность алгоритмов разработанного подхода проверена в ходе анализа смоделированных и экспериментальных данных для систем флуоресцентной спектроскопии. В качестве практической реализации предложенного подхода

Образец цитирования:
Яцков НН, Апанасович ВВ. Разработка комплексного подхода, основанного на методах имитационного моделирования и интеллектуального анализа данных, для исследования систем прикладной флуоресцентной спектроскопии. *Журнал Белорусского государственного университета. Физика.* 2024;1:4–15 (на англ.).
EDN: VJIKKO

For citation:
Yatskou MM, Apanasovich VV. Simulation modelling and data mining approach for the study of applied fluorescence spectroscopy systems. *Journal of the Belarusian State University. Physics.* 2024;1:4–15.
EDN: VJIKKO

Авторы:
Николай Николаевич Яцков – кандидат физико-математических наук, доцент; заведующий кафедрой системного анализа и компьютерного моделирования факультета радиофизики и компьютерных технологий.
Владимир Владимирович Апанасович – доктор физико-математических наук, профессор; профессор кафедры системного анализа и компьютерного моделирования факультета радиофизики и компьютерных технологий.

Authors:
Mikalai M. Yatskou, PhD (physics and mathematics), docent; head of the department of systems analysis and computer simulation, faculty of radiophysics and computer technologies.
yatskou@bsu.by
Vladimir V. Apanasovich, doctor of science (physics and mathematics), full professor; professor at the department of systems analysis and computer simulation, faculty of radiophysics and computer technologies.
apanasovichv@gmail.com
https://orcid.org/0000-0003-4525-4234

создана вычислительная платформа *FluorSimStudio* для обработки измерений флуоресценции с временным разрешением. Цифровая платформа представляет собой открытую систему и позволяет добавлять сложные модели анализа с учетом создания новых алгоритмов моделирования и обработки данных. Применение комплексного подхода повышает эффективность исследования биофизических систем при анализе больших данных.

Ключевые слова: флуоресцентная спектроскопия; обработка данных; имитационное моделирование; интеллектуальный анализ данных; вычислительная платформа.

SIMULATION MODELLING AND DATA MINING APPROACH FOR THE STUDY OF APPLIED FLUORESCENCE SPECTROSCOPY SYSTEMS

***M. M. YATSKOU*[a], *V. V. APANASOVICH*[a]**

[a]*Belarusian State University, 4 Niezaliezhnasci Avenue, Minsk 220030, Belarus*
Corresponding author: M. M. Yatskou (yatskou@bsu.by)

For the study of biomolecular compounds in applied fluorescence spectroscopy supporting systems an integrated approach, based on simulation modelling and data mining methods and including simulation models of physical processes, methods and algorithms for data mining, and software for studying molecular and cellular systems is proposed. The idea of an integrated approach is in using simulation modelling of biophysical processes occurring in the object of study, selecting the most informative experimental data, and determining the characteristics of the object using data mining algorithms. The effectiveness of the algorithms of the proposed approach is verified by analysing simulated and experimental data of fluorescence spectroscopy systems. As a practical implementation of the developed integrated methodology, the computational platform *FluorSimStudio* was developed for processing time-resolved fluorescence measurements. The digital platform is an open system and allows addition of complex analysis models, taking into account the development of new modelling and data processing algorithms. The use of complex analysis improves the efficiency of studying biophysical systems during big data analysis.

Keywords: fluorescence spectroscopy; data processing; simulation modelling; data mining; computational platform.

Introduction

Experimental fluorescence spectroscopy methods are used to investigate the optical properties of molecular compounds and therefore are applied in the studies of artificial photonic materials, protein complexes, biopolymers, biological membranes, cell biomarkers and organic tissues [1–3]. The standing development of methods is driven due to the improvements of effective molecular fluorophores, including genetically expressed proteins (for example, green fluorescent protein (GFP)), semiconductor nanoparticles and quantum dots, optical systems for laser excitation and photon detection, allowing for high-precision measurements, and computer technologies for data storage and processing [4–6]. Novel experimental high-throughput techniques, integrating pulsed, phase and modulation methods for detecting fluorescence decay times, form the basis of modern fluorescence microscopy and allow obtaining big datasets, characterised by high spectral, time and spatial resolution [6; 7]. The main fluorescence spectroscopy and microscopy techniques for studying complex molecular systems in «cuvettes» and living cells are fluorescence-lifetime imaging microscopy, fluorescence recovery after photobleaching and its derivatives (fluorescence loss in photobleaching and fluorescence localisation after photobleaching), fluorescence fluctuation spectroscopy (combining fluorescence correlation spectroscopy, fluorescence cross-correlation spectroscopy, photon counting histogram (PCH) and fluorescence intensity distribution analysis), fluorescence sensing [1; 7].

The advantage of the modern experimental fluorescence spectroscopy methods is the expanding degree of accuracy of measurements, achieved through the use of multichannel spectral, time and spatial resolutions, which significantly enhances the volume of recorded data, but simultaneously increases the efficiency of studying physical processes with a wide dynamic range of changes in parameters and measurement conditions, and allows to study complex multicomponent molecular systems in a series of clinical experiments. One of the major limitations in processing big fluorescence spectroscopy data is the lack of universal effective automated supporting and decision-making systems, including protocols for planning and conducting experiments, software for processing and analysing data, modelling and interpreting the studied biophysical processes. Presently,

systems for collecting and processing fluorescence spectroscopy data have been implemented to study optical processes in molecular compounds [6; 8]. Their advantages are acceptable performance and barely reasonable support of experimental investigations. However, these systems are not complex, used for specific types of biomolecular compounds and specialised experimental equipment, they are not organised in the form of a universal technique or approach to big data analysis, and provide a limited set of physical interpretation models and software analysis tools, which lets us to conclude that it is necessary to systematise existing solutions, integrating the most effective methods of data analysis, physical models (for example, based on simulation modelling) into a complex approach. The development of new improved supporting and decision-making systems should simplify and automate the processing of fluorescent experimental measurements, increase the accuracy of estimated parameters, and expand the limits of interpretation and prediction of physical processes.

The existing data analysis approaches to processing fluorescence spectroscopy data can be divided into classical and modern, based on machine learning or data mining algorithms. Classical methods consider separate or joint analysis of datasets using deconvolution, least squares, maximum likelihood, Bayesian, target and global analysis to estimate the parameters of mathematical models of optical processes and systems [9]. New approaches are based on: i) projection transformations and following parameter estimation (for example, transformation of fluorescence intensities into the phasor space (phasor analysis)); ii) using machine learning techniques, mainly artificial neural networks and ensemble algorithms, to estimate the model parameters; iii) segmentation of cell or tissue images and subsequent classification by a machine learning algorithm [10–14]. The main disadvantages of existing data processing methods are limited efficiency, that is due to the use of nonphysical analytical models (multi-exponential or polynomial decompositions), poor accuracy in parameter estimating when analysing noisy data (phasor analysis, neural networks), slow computations (global and Bayesian analysis), the need for the large training datasets (neural networks), special requirements for computing resources (the usage of video cards or multiprocessor nodes to accelerate neural network computing), and finally the lack of specialised software for automated data processing. Therefore, the primary task is to develop an integrated data analysis approach that eliminates the main drawbacks of existing methods, which would include physical models of the processes and systems under study, effective methods and software for processing a series of fluorescence spectroscopy data [15–17]. It should include descriptions of molecular systems at various levels of generalisation, from simple molecular compounds in solutions and films to complex cellular systems involved in various diseases.

In this paper, we propose an integrated approach based on simulation modelling and data mining methods for the study of biomolecular compounds in applied fluorescence spectroscopy systems. It includes simulation models of physical processes, methods and algorithms for data mining, and software for automating data analysis. As a practical implementation, the developed integrated methodology is integrated into the computational platform *FluorSimStudio* for processing fluorescence kinetic curves obtained through time-resolved fluorescence experiments.

Methodology

It is assumed that through a series of experiments generating big datasets some object, i. e. biophysical process or biomolecular compound, is investigated, whose essential properties or characteristics, for example, a set of biophysical parameters $A = \{a_1, a_2, \ldots, a_P\}$ (e. g., electronic excitation energy transfer rate constants, protein concentrations and diffusion coefficients, biochemical reaction rate constants, etc.), must be determined during data analysis. The number of parameters P is determined depending on the level of detail (abstraction) of physical description and complexity of the investigated object. The parameters should be sufficient to adequately explain the behaviour of the object. In an integrated approach data mining algorithms are applied to multidimensional datasets to select the most informative or significant data for further in-depth analysis and finding estimates of parameters A using simulation models. Let us consider the main components of a complex approach.

Data. Let there be N observations, samples or measurements $H_E = \{E_1, E_2, \ldots, E_N\}$ of the investigated object E, kept in the database or obtained in a real physical experiment. Let us consider the formalisation of conducting an experiment to investigate the object under study (fig. 1).

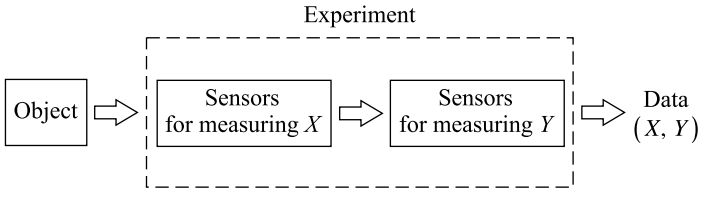

Fig. 1. Scheme of studying an object in a physical experiment

We assume that the experiment combines a set of sensors to record some properties of the observations of the object. The registered object properties are known as features, attributes, or variables and can be of two types. The first ones, denoted X, are independent measurements, including external influences (this can also include time as a signal), set by the experiment designer or researcher. The second ones, denoted Y, are measurements depending on the selected values of the characteristics of the first group. Independent measurements X are usually called features (or inputs, predictors, independent variables), and dependent Y – target variables (or outputs, responses, dependent variables) that determine solutions to data analysis problems. As a result of measuring the properties of observations, a dataset $D = \{X, Y\}$ (or (X, Y)) is recorded that combines independent and dependent variables of observations. The data structure of the object is considered in terms of selected attributes X and dependent variables Y.

Measurements over observations form feature vectors $X_1, X_2, \ldots, X_K$ and an input data matrix X. Measurements over observations with fixed values $X_1, X_2, \ldots, X_K$ form vectors of target features or output characteristics $Y_1, Y_2, \ldots, Y_R$ and an output data matrix Y. The data (X, Y) obtained from a real experiment are presented in the form of a table, containing the recorded values of the properties of the object.

A set of multidimensional data recorded in some physical experiment and representing observations or measurements H_E of the investigated object E

H_E	X				Y			
	X_1	X_2	…	X_K	Y_1	Y_2	…	Y_R
E_1	x_{11}	x_{12}	…	x_{1K}	y_{11}	y_{12}	…	y_{1R}
E_2	x_{21}	x_{22}	…	x_{2K}	y_{21}	y_{22}	…	y_{2R}
…	…	…	…	…	…	…	…	…
E_N	x_{N1}	x_{N2}	…	x_{NK}	y_{N1}	y_{N2}	…	y_{NR}

The model of the data can be represented as

$$Y = \Xi(X), \tag{1}$$

where Ξ is a set of correspondence operators that transforms independent features X into dependent variables Y.

Data mining. Let us consider the formulation of the data analysis problem. The general task of data analysis is to find functional transformations Π (correspondence estimates Ξ) and their parameters Θ, transforming the original set of features X into a theoretical set of dependent characteristics

$$Y^T = \Pi(X, \Theta), \tag{2}$$

such that the condition of minimal difference between the observed output characteristics Y and the theoretical Y^T is satisfied.

By data mining models we assume mathematical models that provide solutions to problems (1) and (2), i. e. functional transformations (for parametric models) $\mathcal{M} = \{\Pi(X, \Theta) | \Theta \in D_\Theta\}$, where $\Pi: X \times \Theta \to Y$, D_Θ – set of admissible values of parameters Θ or parameter space. The model parameters Θ, in the general case, do not coincide with the physical parameters A.

In practice, such models are built based on existing known experimental data, representing the so-called expert-labeled, reference or training data. The specificity of data mining models is the lack of consideration of the physical principles of the processes in the object under study. The model is built according to so-called precedents or existing examples, based on behavioural assessments of which it is planned to predict the behaviour of the object in the future. The most popular data mining models are cluster algorithms, decision trees, associative rules, mathematical functions (analytical models), and neural networks [18].

As data mining methods, we consider computational algorithms μ for finding estimates of unknown parameters Θ of the data mining models, i. e. the method is a mapping $\mu: X \times Y \to \mathcal{M}$, which associates an arbitrary finite data sample (X, Y) with some algorithm for determining parameters Θ, such that the condition of minimal difference between the observed output characteristics Y and the theoretical Y^T is satisfied. The most popular data mining methods are statistical analysis (regression and variance analysis, data dimensionality reduction), classification and prediction (artificial neural networks, decision trees, k-nearest neighbours, Bayesian networks), cluster analysis, optimisation, association rule search and data visualisation [18].

The experimental data mining diagram is shown in fig. 2. The overall task of data analysis can be divided into components that form the basis of data mining. The main tasks of data mining include classification, regression, cluster analysis, data dimensionality reduction, association rule search and visualisation [18].

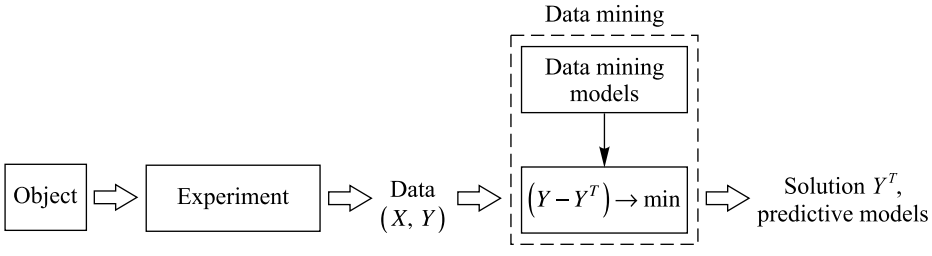

Fig. 2. Scheme of studying an object in an experiment using data mining

An important feature of data models is that they can be used to build a predictive model and obtain an exact solution to the problem. This is often sufficient, for example, for classifying images of cancer cells or multi-exponential smoothing of fluorescence decay kinetic curves [19–21]. The disadvantage of this approach is the impossibility of reproducing a detailed description of the physical processes occurring in the object. To eliminate this drawback, it is necessary to use physical models, for example, based on simulation modelling.

Simulation modelling. The basis of simulation modelling is Monte Carlo methods [22–24], which are stochastic modelling algorithms based on the use of random numbers and statistical probabilities for solving applied problems. Traditionally, Monte Carlo methods find applications in two directions: when checking the reliability of approximate solutions obtained as a result of analytical calculations, i. e. to confirm the developed theories by numerical experiment [15; 25]; to compare simulated and experimental data, followed by a deeper interpretation of the data in terms of simulation models [26; 27]. The presented work deals with the second direction of applications of simulation modelling methods.

In modelling, the investigated object E can be considered as a biophysical process (for example, electronic excitation energy transport in a molecular system or actin polymerisation in a cell) or a biomolecular system (molecules, cells or cell populations). Let object E be described by a mathematical (analytical or simulation) model M (fig. 3).

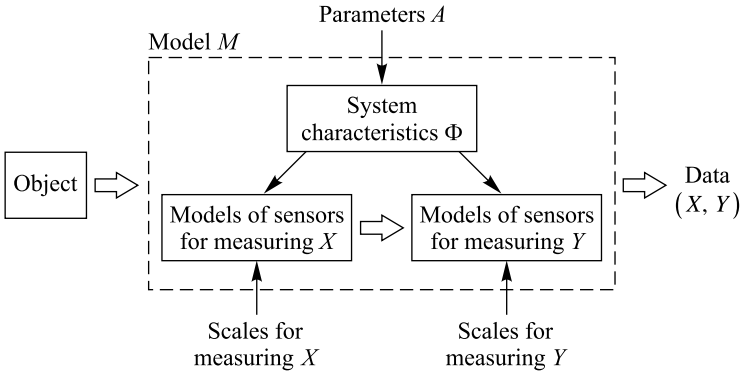

Fig. 3. Scheme of the mathematical model of the object

When constructing a mathematical model, it is necessary to take into account the measurement scales of the features X and the characteristics of the system states Φ (systems of equations or modelling algorithms that describe the behaviour of the object) for given values of the physical parameters of the model A. The characteristics of the system states Φ are represented by a matrix of features X, including response components system or output signals Y. Then the mathematical model of the object can be represented as the expression

$$M = \{X, \Phi, A, F\},$$

where F is the operator of functional transformations

$$Y = F\{X, A\}.$$

The mathematical model describes the real processes occurring in the investigated object. It can be presented in the form of a «white box», since it takes into account the physics of the occurring processes in the object. The dataset for subsequent analysis is a table of characteristics of system states X and output

signals Y. At the initial moment of the investigation, the internal structure and relationships between the components of the object E are known. The task is to clarify the functional dependencies F and estimate the model parameters A. It should be noted that the «black box» models $\mathcal{M}$ in the form of data mining models are analytical approximation models ($F = \Pi$, parameters Θ), not taking into account the physics of the process under study.

An integrated approach to big data analysis. This work suggests to use simulation modelling and data mining methods to study biomolecular systems, a feature of which is the use of simulation modelling algorithms to reproduce biophysical processes in the systems under study. The idea of an integrated approach consists of studying the investigated object using simulation modelling of biophysical processes occurring in it, comparing simulated and the most informative experimental data, selected by the data dimensionality reduction methods, and determining the parameters of physical processes using optimisation algorithms.

A diagram of the studying an object (some biophysical process or biomolecular compound) using the developed integrated approach is shown in fig. 4.

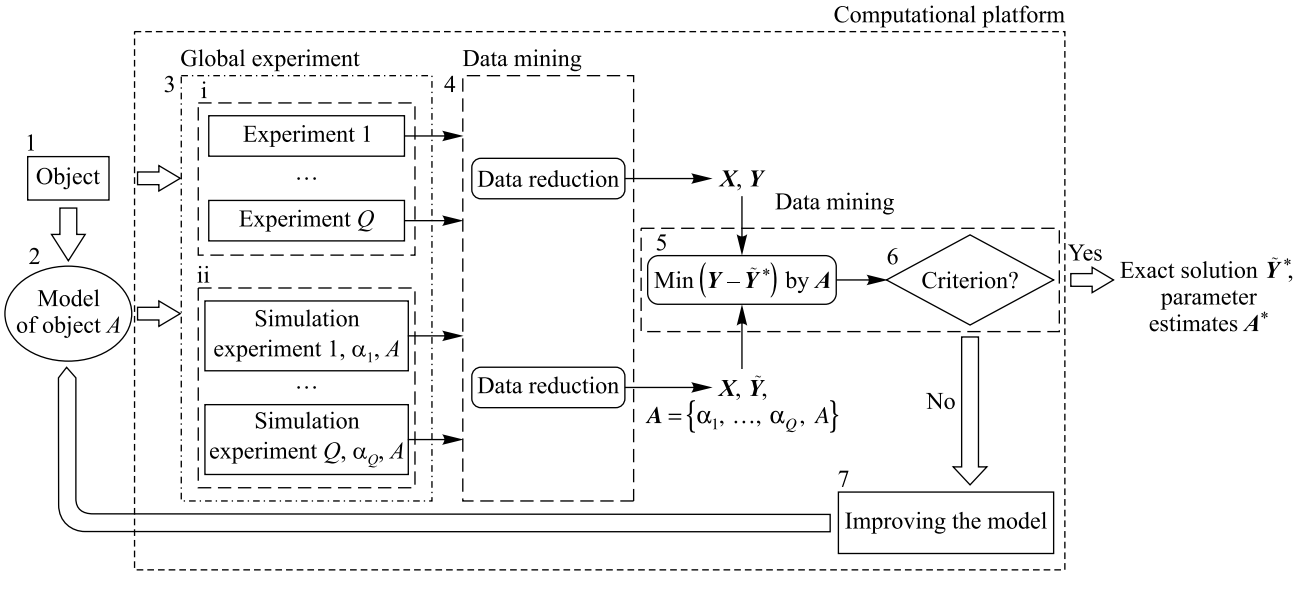

Fig. 4. Scheme of studying an object using the developed integrated approach

The study of the object (block 1) is carried out by considering the physical model of the object (block 2) and a series of Q real and simulated experiments that form a global experiment (block 3). Simulation experiment includes modelling of the experimental equipment and the object. The data of individual experiments are converted to a single format in order to reduce and eliminate inhomogeneities of various distortions associated with specific experiments. Filtering, normalisation, vectorisation, or special data transformations, such as logarithmic, are performed to reduce the effect of outliers. In block 4, data dimensionality reduction is performed. Data are compressed by the dimensionality reduction methods in order to exclude uninformative, redundant data or noise, and essential informative data are extracted. Sets of transformed data from various experiments are collected into a combined dataset $(\boldsymbol{X}, \boldsymbol{Y})$ for subsequent processing using data mining and simulation modelling methods (block 5). The parameters of individual experiments $\alpha_1, \ldots, \alpha_Q$, and the model of object A are collected into a single set $\boldsymbol{A}$ and then refined during the analysis of the combined data. The analysis of individual experiments can be carried out independently or in a complex manner. The advantage of integrated analysis is the combination of data from various experiments into one large set, which provides a generalisation and an increase in the statistical power of the results and, as a consequence, an increase in the accuracy of the analysis. Some parameters $\boldsymbol{A}$ are fixed (they are global for the experiments), limited (in the case of dependent experiments), or remain free for accurate estimation using optimisation methods. Optimisation methods are used to evaluate free or adjustable parameters $\boldsymbol{A}^*$ of the global simulation model of the object, built on the basis of models of individual experiments. If the desired accuracy of the correspondence between the experimental and simulated data is achieved, which is determined by the given statistical criterion in block 6 (for example, the quantitative χ^2, Kolmogorov – Smirnov and Romanovskii criteria, as well as diagram plots of the weighted residuals, their autocorrelation functions, and histograms [15]), then the analysis process is completed and a solution is provided (an estimated set of parameters and an accurate mathematical model of the object capable of

predicting its behaviour in the pre-cases of a desirable accuracy). Otherwise, in block 7, the description of the object is improved (including the deepening of the object formalisation, collection of new data, changing models, conducting additional experiments, changing the parameters of the object or environment) and move to block 2 to perform the next iteration of data analysis. The presented scheme is a general approach, the specific implementation is determined by the type of the solved problem and should be designed taking into account the specifics of the conducted experiments.

In the integrated approach, various methods and algorithms should be tried, moving from simple models to complex ones, performing a gradual complication of the models and increasing the circumstantiation of the process under consideration. When working with big data, it is necessary to choose an adequate level of refrainment of the mathematical model, corresponding to the desired depth of the object investigation, the volume of datasets and the power of the available computing resources.

Error analysis. When carrying out computer modelling, it is important to confirm the adequacy of the simulation models, the reliability, confidence and reproducibility of the results obtained, as well as achieving the desired modelling accuracy. Assuming that the optimal model parameters have been chosen, three stages of confirming its adequacy can be considered, i. e. the model must meet the following requirements [28–30]:

• be physical, based on consideration of the physics of processes «from first principles», which includes the selection of the most accurate laws for describing probability distributions for the random variables and processes under consideration;

• be confirmed by an analytical description or experiment under certain control conditions, various balance equations and internal technical verification tests of algorithms. In the literature, this requirement is often referred to as model verification;

• be confirmed experimentally, for example, to ensure a minimum error when compared with experimental data according to some pre-established statistical criterion. Provided that the first and second requirements are met, the criterion allows to assess the validity or adequacy of the models. The following test results are possible:

– the criterion value is unacceptable to confirm the statistical significance of the model. Therefore, the model is not supported by experimental data and is assumed to be inadequate. Inadequacy may be a consequence of suboptimal selection of model parameters or model inaccuracy. Inadequacy should be eliminated by clarifying the model parameters, expanding the formalisation of the modelled processes, moving to a deeper level of the consideration, or completely replacing the model, achieving an improvement in the value of the criterion;

– the criterion takes optimal values for a large set of models, which corresponds to the limitations of the selected criterion or the redundancy of models (for example, in problems of analysing fluorescence decay curves, this is a large number of exponentials, a high-degree polynomial, a multilayer neural network, a simulation model taking into account the modelling of insignificant processes, not affecting the output characteristic);

– the criterion tends to the best value in an extreme or excessively accurate manner, which corresponds to overtraining of the model. The situation is typical for overfitting a regression model when approximating experimental noise in the data – the model loses an important generalisation property.

The most reliable model is assessed using cross-validation or bootstrap algorithms [31; 32].

Examples of biophysical systems for investigation in integrated data analysis approach. Let us consider the possibilities of applying an integrated approach to the analysis of big data using examples of molecules, biopolymers, proteins and cell systems, studied by fluorescence spectroscopy methods and representing groups of data associated with the processes of electronic excitation energy transfer at the level of molecules and their compounds and with the processes of protein interaction at the cell level. These systems and processes are studied in the construction of molecular photonic artificial antenna systems and in the diagnosis of oncological diseases; they are combined by such an area of experimental research as applied fluorescence spectroscopy. In the experiments, the fluorescence of molecular compounds or luminescent dyes, that mark the molecules of the samples, is studied. The samples are exposed to laser radiation at the excitation wavelength of the molecules or dyes, and the intensity of fluorescence emission is then recorded. Optical processes and molecular systems are studied using the intensity of emitted fluorescence. What is common in the study of objects of these systems are the area of experimental methods for obtaining data, formalisation of the description of objects and their observations, analysis and modelling algorithms, mathematical models used to describe biophysical processes, formalism of data presentation (see table). We adhere to the following scheme for describing the system: object of study – observations or measurements of the object – features or attributes of the object – formulation of the problem to be solved in terms of an integrated approach.

Molecules. The object of study is the process of energy transfer in artificially created molecular systems [1; 33]. Object observations are the optical spectra and fluorescence decay curves recorded for the molecular compounds

of interest under specified experimental conditions. Features are absorption (emission) wavelengths, time counts of photon registration on the detector, physical parameters characterising experimental samples (concentration of molecules, type of solution, temperature, day of measurement, etc.). The dependent variable is the fluorescence intensity at specified times, excitation and emission wavelengths. The problem of regression is solved, namely, finding a model and its parameters that satisfactorily describe the kinetic curves of fluorescence decay. Simulation modelling of electronic excitation energy transfer processes in molecular systems is carried out. In an integrated approach, fluorescence decay curves are grouped in clusters, the medoids of clusters are found, the medoid curves are analysed using simulation models, the result of which is the estimated parameters of energy transfer processes. Different clusters of decay curves are associated with molecular compounds based on the estimated parameters. In the case of a small set of fluorescence decay curves, the integrated approach is reduced to a single analysis of the decay curves using simulation models according to the simulation-based approach [34].

Proteins. The object of study is the processes of diffusion and aggregation of proteins in various environments (for example, proteins involved in the formation of cancer cells [35; 36]). In fluorescence fluctuation spectroscopy experiments, the fluorescence intensity of molecular complexes is recorded, which allows to estimate the size of the protein complex [37]. Observations of the object are PCHs of fluorescence intensity fluctuations for fluorescently-labeled proteins at a given recording time interval. Features are histogram channels represented by the photon frequencies detected during a certain short time interval. The dependent variable is a marker of the type of protein or molecular complex. Simulation modelling of the proteins diffusion and aggregation processes is carried out. In the integrated approach, the grouping of PCHs, the determination of medoids of PCH clusters and the analysis of PCHs using simulation models are performed, the result of which is the estimated parameters of the proteins diffusion and aggregation processes.

In fluorescence spectroscopy experiments conducted on cells (*in vitro* and *in vivo*), fluorescence analysis allows to estimate the parameters of association and dissociation reactions (or in the general case of polymerisation) to molecular protein complexes [38; 39]. Observations are kinetic curves of increase and recovery of fluorescence after photobleaching of proteins. Features are the time of occurred and detected fluorescence emission events of a luminescent protein. The dependent variable is protein fluorescence recorded at the indicated times. The regression problem is solved, namely, finding a model of protein polymerisation and its parameters that satisfactorily describe the fluorescence intensity. Simulation modelling of protein polymerisation processes in a cell is carried out. In the integrated approach, fluorescence curves are processed, data cluster medoids are found and analysed using simulation models, which result in estimated parameters of protein polymerisation processes.

Cells. The object of the study is a cancer disease, determined by the characteristics of cell microobjects in a luminescent image [40; 41]. Observations are highlighted (segmented) cells in the image (contours of nuclei or cytoplasms), affected by disease. Features (properties of segmented microobjects in the image) are cell characteristics obtained as a result of segmentation of cell contours (size, colour, orientation, etc. [19]). The dependent variable is a marker of cell cancerous type or stage of disease. The problem of classifying cells into cancerous (non-cancerous) or determining the stage of the disease is solved. Simulation modelling reproduces the stages of the disease with given parameters. In an integrated approach, the stages of cancer are modelled, parameters corresponding to cancerous (non-cancerous) cells are determined, microobjects of images are classified into one or another type of cell susceptible to a certain stage of the disease. Simulation modelling may not be used if expert solutions are given in the form of labeled images of cancer cells.

Computational platform for the implementation of integrated approach methods and algorithms. The developed integrated approach to the analysis of big experimental data can be practically implemented in the form of a computing platform or programming environment that combines computational algorithms of mathematical models and analysis methods, as well as auxiliary software tools for data processing. In the structure of the computational approach, the platform integrates implementations of simulation models, analysis algorithms and analysis quality assessment (see fig. 4).

As a demonstrative example of the practical implementation of the developed complex methodology, integrating simulation modelling and data mining algorithms, the computational platform *FluorSimStudio* has been developed for processing fluorescence kinetic curves in time-resolved fluorescence experiments [42]. The user's work is carried out through a web application hosted at *https://dsa-cm.shinyapps.io/FluorSimStudio*. An example of the package interface window is shown in fig. 5. The main interface window consists of nine panels corresponding to six stages of analysis: loading, modelling and clustering data, reducing data dimensionality by the principal component analysis, fitting medoids (data analysis), visualising and interpreting the results, information about the authors of the development, and instructions for using the computational resource.

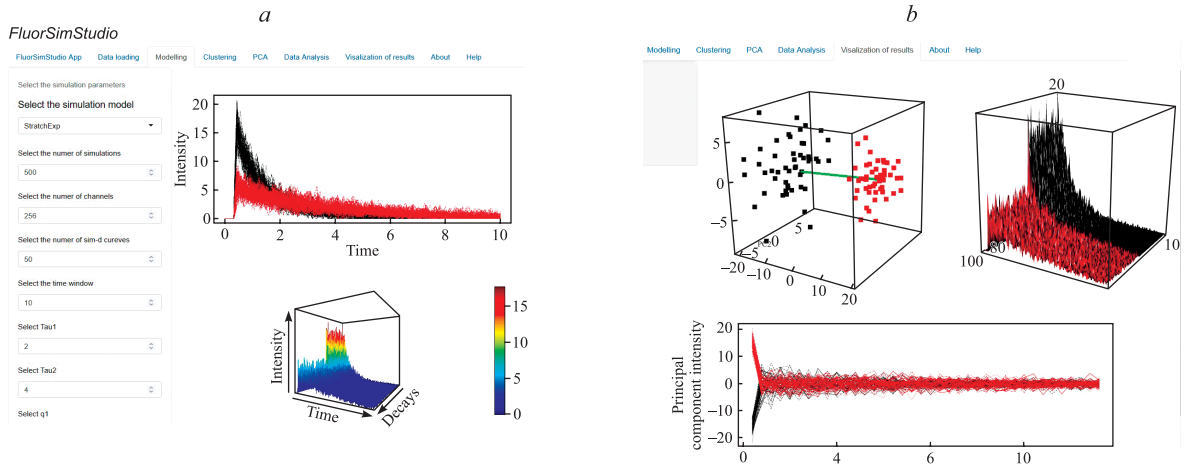

Fig. 5. FluorSimStudio web application interface windows:
examples of simulation modelling the stretched-exponential fluorescence decay dataset (*a*)
and analysis results visualisation (*b*) of clustered fluorescence decay curves

The platform is not intended for data collection and storage, development of new analysis algorithms, organisation of parallelised computational systems for big data analysis. Its main task is to provide the user with computing tools for the implementation of developed simulation models, methods of modelling and analysis of big data, instruments for assessing the quality of analysis, visualisation and interpretation of data.

Results

Application of an integrated approach to the analysis of molecular systems. We consider two examples that confirm the fundamental applicability of the developed integrated approach to the analysis of molecular systems in fluorescence spectroscopy experiments.

Time-resolved fluorescence spectroscopy. The effectiveness of the algorithms of the developed approach was tested during the analysis of fluorophore systems under various parameters of a computational experiment. Simulated time-resolved fluorescence datasets, representing three fluorophore systems and characterised by one-, two- and stretched-exponential fluorescence decay laws were studied [16]. The stretched-exponential model is a represent of the donor fluorescence intensity decay in the presence of Förster type of energy transfer in a donor – acceptor molecular system. The donor fluorescence in three dimensional space can be written as

$$i\left(t,\, i_0,\, q,\, \tau_D\right) = i_0 \exp\left\{-\frac{t}{\tau_D} - q\left(\frac{t}{\tau_D}\right)^{\frac{1}{2}}\right\},$$

where i_0 is the fluorescence intensity at time $t = 0$; $q = 0.5\left[C_A\right]/\left[C_{A0}\right]$, where C_{A0} and C_A are the critical and actual concentrations of acceptors; τ_D is the fluorescence decay time of donors [16]. The application of the algorithms of the developed approach to the analysis of datasets made it possible to accurately determine the fluorescence lifetimes of fluorophores. The accuracy of the estimated parameters by complex analysis is higher than in the case of using the classical approach based on separate processing of each dataset using analytical models of fluorescence decay (fig. 6, *a*). The developed approach requires significantly less time and the number of calculations of the theoretical model, and allows faster and more accurate determination of the parameters of biophysical and optical processes in molecular compounds in comparison with the classical method. The use of simulation models of optical-physical processes significantly increases the efficiency of parameter estimation in the case of analysis of complex molecular systems, such as photonic antennas based on metalloporphyrin films or zeolite crystals [43; 44], when the parameters of the molecular environment and the mechanisms of electronic excitation energy transfer, necessary to create accurate analytical models, are unknown.

Fluorescence fluctuation spectroscopy. The effectiveness of the analysis algorithms developed within the framework of the proposed approach was confirmed by simulated and experimental PCH measurements of fluctuations in the fluorescence intensity of the green fluorescent protein GFP-S65T [45; 46]. Analysis of experimental data on the GFPs in cell lysates revealed the presence of monomeric and dimeric forms of proteins (fig. 6, *b*–*f*). Monomers of the GFP demonstrate a spherical data cluster in the space of the first two principal

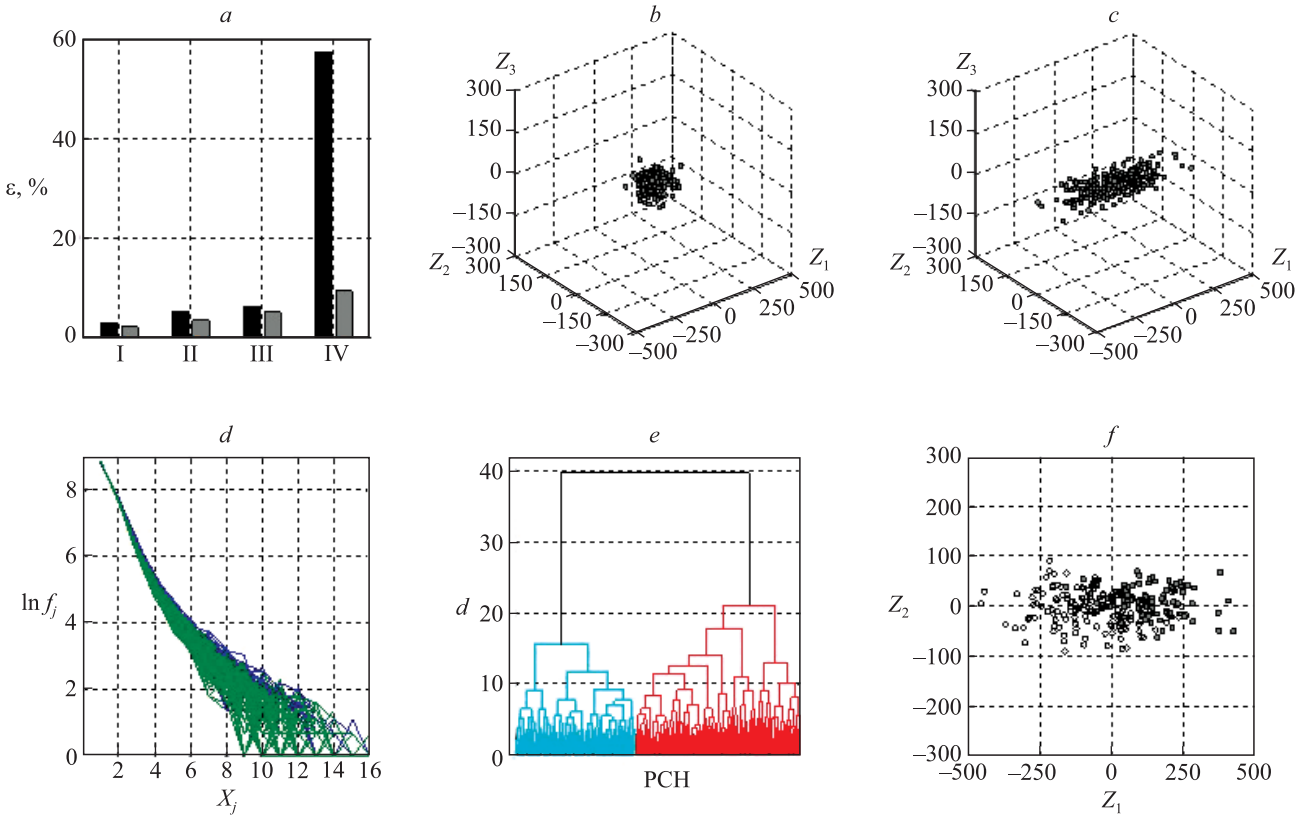

Fig. 6. Examples of the application of an integrated approach to the analysis of molecular compounds in experiments with time-resolved fluorescence (*a*) and fluorescence fluctuation (*b–f*) spectroscopies: (*a*) error ε in assessing the accuracy of reconstructing the parameters of simulated fluorescence decay curves of fluorophore systems characterised by one-exponential (I and II), two-exponential (III) and stretched-exponential (IV) laws of fluorescence decay, using classical (black) and developed (gray) methods. Digital labels (I–IV) of the abscissa axis denote the following modelling parameters: fluorescence decay times $\tau_1 = 2$ ns and $\tau_2 = 4$ ns, number of curves is 200, standard deviation of parameters $\sigma = 0.1$ (I); $\tau_1 = 1.4$ ns and $\tau_2 = 2$ ns, number of curves is 200, $\sigma = 0.1$ (II); $\tau_1 = 0.5$ ns and $\tau_2 = 2$ ns, their contributions (normalised to one) $p_1 = 0.2$, $p_2 = 0.8$ and $p_1 = 0.8$, $p_2 = 0.2$, respectively, for two sets of decay curves, number of curves is 200, $\sigma = 0.1$ (III); donor fluorescence decay time $\tau_D = 2$ ns, acceptor concentration q is equal 1 and 0.2 for two sets of decay curves, number of curves is 200, $\sigma = 0.1$ (IV); (*b*) and (*c*) are photon counting histograms PCHs of monomeric and dimeric forms of the green fluorescent protein GFP in the coordinates of the most informative components Z_1, Z_2 and Z_3, calculated by the principal component method; (*d*) PCH on a logarithmic scale in the space of the initial features $X_1, X_2, \ldots, X_{16}$, f_j are frequencies of occurrence of the number of photons during a certain short time interval; (*e*) dendrograms of PCH, *d* is a measure of similarity of clusters; (*f*) PCH in the space of the first two principal components Z_1 and Z_2. In illustrations *d–f* colours and symbols indicate monomeric and dimeric forms of proteins

components (see fig. 6, *b*), while an elongated ellipsoidal cloud is observed for a mixture of monomeric and dimeric forms of protein compounds (see fig. 6, *c*). Note that accurate separation of monomeric and dimeric forms of a protein is difficult to achieve using classical analytical methods, which involve a separate analysis of PCHs. Further assessment of the parameters of protein complexes can be made by analysing the medoids of the resulting PCH clusters using classical analysis or simulation algorithms.

Conclusions

The integrated approach to processing big datasets in applied fluorescence spectroscopy has been developed, which is based on simulation modelling and data mining methods for the study of optical processes in biophysical systems. Its main feature is the use of simulation modelling algorithms to reproduce biophysical processes in the systems under study and data mining to determine the most informative data. The effectiveness of the algorithms was verified by analysing simulated and experimental data representing systems of molecules and proteins that are studied in time-resolved fluorescence and fluorescence fluctuation spectroscopy experiments. The developed analysis approach, in comparison with the classical one, quickly and more accurately determines the parameters of biophysical and optical processes in molecular compounds. The proposed methodology of the developed integrated approach was realised in the computational platform *FluorSimStudio*, intended for processing fluorescence decay curves of molecular systems. It provides high productivity of processing

large fluorescence datasets, is hosted on the server and can be used in the educational process and for the study of time-resolved fluorescence spectroscopy systems. By the developed integrated approach, it is possible to increase the accuracy of assessing the studied characteristics of biophysical processes in comparison with the classical approach, based on separate processing of each dataset, to deepen knowledge about the physics and essence of the processes under study, to create new predictive tools when analytical models do not exist or the derivation of analytical solutions is difficult due to increasing complexity of a system represented by big data.

References

1. Lakowicz JR. *Principles of fluorescence spectroscopy.* 3rd edition. New York: Springer; 2006. XXVI, 954 p. DOI: 10.1007/978-0-387-46312-4.

2. Verveer PJ, editor. *Advanced fluorescence microscopy: methods and protocols.* New York: Humana Press; 2015. XI, 294 p. (Methods in molecular biology; volume 1251). DOI: 10.1007/978-1-4939-2080-8.

3. Demchenko AP. *Introduction to fluorescence sensing. Volume 1, Materials and devices.* 3rd edition. Cham: Springer; 2020. XXII, 657 p. DOI: 10.1007/978-3-030-60155-3.

4. Weinacht TC, Pearson BJ. *Time-resolved spectroscopy: an experimental perspective.* Boca Raton: CRC Press; 2019. XII, 358 p. (Textbook series in physical sciences). DOI: 10.1201/9780429440823.

5. Buckup T, Léonard J, editors. *Multidimensional time-resolved spectroscopy.* Cham: Springer; 2019. IX, 320 p. (Topics in current chemistry collections). DOI: 10.1007/978-3-030-02478-9.

6. Gryczynski Z, Gryczynski I. *Practical fluorescence spectroscopy.* Boca Raton: CRC Press; 2020. XX, 772 p. DOI: 10.1201/9781315374758.

7. Cox G, editor. *Fundamentals of fluorescence imaging.* Singapore: Jenny Stanford Publishing; 2019. XVIII, 458 p. DOI: 10.1201/9781351129404.

8. König K, editor. *Multiphoton microscopy and fluorescence lifetime imaging: applications in biology and medicine.* Berlin: Walter de Gruyter; 2018. XXX, 450 p. DOI: 10.1515/9783110429985.

9. Datta R, Heaster TM, Sharick JT, Gillette AA, Skala MC. Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications. *Journal of Biomedical Optics.* 2020;25(7):071203. DOI: 10.1117/1.JBO.25.7.071203.

10. Datta R, Gillette A, Stefely M, Skala MC. Recent innovations in fluorescence lifetime imaging microscopy for biology and medicine. *Journal of Biomedical Optics.* 2021;26(7):070603. DOI: 10.1117/1.JBO.26.7.070603.

11. Xiao D, Zang Z, Xie W, Sapermsap N, Chen Y, Li DDU. Spatial resolution improved fluorescence lifetime imaging via deep learning. *Optics Express.* 2022;30(7):11479–11494. DOI: 10.1364/OE.451215.

12. Wang Q, Li Y, Xiao D, Zang Z, Jiao Z, Chen Y, et al. Simple and robust deep learning approach for fast fluorescence lifetime imaging. *Sensors.* 2022;22(19):7293. DOI: 10.3390/s22197293.

13. Ji M, Zhong J, Xue R, Su W, Kong Y, Fei Y, et al. Early detection of cervical cancer by fluorescence lifetime imaging microscopy combined with unsupervised machine learning. *International Journal of Molecular Sciences.* 2022;23(19):11476. DOI: 10.3390/ijms231911476.

14. Place BC, Troublefield CA, Murphy RD, Sinai AP, Patwardhan AR. Machine learning based classification of mitochondrial morphologies from fluorescence microscopy images of *Toxoplasma gondii* cysts. *PLOS One.* 2023;18(2):e0280746. DOI: 10.1371/journal.pone.0280746.

15. Yatskou MM. *Computer simulation of energy relaxation and transport in organized porphyrin systems* [dissertation on the Internet]. Wageningen: Ponsen & Looijen Printing Establishment; 2001 [cited 2023 April 4]. 176 p. Available from: https://edepot.wur.nl/193545.

16. Yatskou MM, Skakun VV, Apanasovich VV. Method for processing fluorescence decay kinetic curves using data mining algorithms. *Journal of Applied Spectroscopy.* 2020;87(2):333–344. DOI: 10.1007/s10812-020-01004-3.

17. Yatskou MM, Apanasovich VV. Simulation modelling and machine learning platform for processing fluorescence spectroscopy data. In: Tuzikov AV, Belotserkovsky AM, Lukashevich MM, editors. *Pattern recognition and information processing. Revised selected papers of the 15th International conference; 2021 September 21–24; Minsk, Belarus.* Cham: Springer; 2022. p. 178–190 (Communications in computer and information science; volume 1562). DOI: 10.1007/978-3-030-98883-8_13.

18. Murphy KP. *Probabilistic machine learning: an introduction* [Internet]. Cambridge: MIT Press; 2022 [cited 2022 July 7]. 864 p. Available from: https://mitpress.mit.edu/9780262369305/probabilistic-machine-learning.

19. Lisitsa Y, Yatskou M, Skakun V, Apanasovich V. Classification methods for the analysis of segmented objects on fluorescent images of cancer cells. *Vestnik of Polotsk State University. Part C, Fundamental Sciences* [Internet]. 2020 [cited 2023 November 1];4:15–22. Available from: https://journals.psu.by/fundamental/article/view/440. Russian.

20. Apanasovich VV, Novikov EG, Yatskov NN, Koehorst RBM, Schaafsma TJ, van Hoek A. Study of the Zn-porphyrin structure by fluorescence spectroscopy methods. *Journal of Applied Spectroscopy.* 1999;66(4):613–616. DOI: 10.1007/BF02675396.

21. Meyer M, Yatskou MM, Pfenninger M, Huber S, Calzaferri G, Digris A, et al. Excitation energy migration in a photonic dye-zeolite antenna: computational techniques. *Journal of Computational Methods in Science and Engineering.* 2003;3(3):395–402. DOI: 10.3233/JCM-2003-3303.

22. Dimov IT. *Monte Carlo methods for applied scientists.* Singapore: World Scientific Publishing; 2008. XV, 291 p. DOI: 10.1142/2813.

23. Rubinstein RY, Kroese DP. *Simulation and the Monte Carlo method.* 3rd edition. Hoboken: John Wiley & Sons; 2017. XVII, 414 p. (Wiley series in probability and statistics). DOI: 10.1002/9781118631980.

24. Binder K, Heermann DW. *Monte Carlo simulation in statistical physics: an introduction.* 6th edition. Cham: Springer; 2019. XVII, 258 p. (Graduate texts in physics). DOI: 10.1007/978-3-030-10758-1.

25. Halavatyi AA, Nazarov PV, Al Tanoury Z, Apanasovich VV, Yatskou MM, Friederich E. A mathematical model of actin filament turnover for fitting FRAP data. *European Biophysics Journal.* 2010;39(4):669–677. DOI: 10.1007/s00249-009-0558-2.

26. Apanasovich VV, Novikov EG, Yatskov NN. Data analysis in time-resolved fluorescence spectroscopy using computer simulation. In: Thompson RB, editor. *Advances in fluorescence sensing technology III. Proceedings of BIOS'97, part of Photonics West; 1997 February 8–14; San Jose, USA.* Bellingham: SPIE; 1997. p. 515–522 (Proceedings of SPIE; volume 2980). DOI: 10.1117/12.273558.

27. Apanasovich VV, Novikov EG, Yatskov NN. Analysis of the decay kinetics of fluorescence of complex molecular systems using the Monte Carlo method. *Journal of Applied Spectroscopy.* 2000;67(5):842–851. DOI: 10.1023/A:1004111716211.

28. Durán JM. *Computer simulations in science and engineering: concepts, practices, perspectives.* Cham: Springer; 2018. XXI, 209 p. (The frontiers collection). DOI: 10.1007/978-3-319-90882-3.

29. Zeigler BP, Muzy A, Kofman E. *Theory of modeling and simulation: discrete event and iterative system computational foundations.* 3rd edition. [S. l.]: Academic Press; 2018. XXIV, 667 p. DOI: 10.1016/C2016-0-03987-6.

30. Tang J, Leu G, Abbass HA. *Simulation and computational red teaming for problem solving.* Hoboken: John Wiley & Sons; 2020. XXVII, 464 p. (IEEE Press series on computational intelligence). DOI: 10.1002/9781119527183.

31. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *IJCAI'95. Proceedings of the 14th International joint conference on artificial intelligence; 1995 August 20–25; Montreal, Canada. Volume 2.* San Francisco: Morgan Kaufmann Publishers; 1995. p. 1137–1143.

32. Liu L, Özsu MT, editors. *Encyclopedia of database systems.* New York: Springer; 2009. CCCXLV, 4355 p. DOI: 10.1007/978-0-387-39940-9.

33. Yatskou MM, Donker H, Koehorst RBM, van Hoek A, Schaafsma TJ. A study of energy transfer processes in zinc-porphyrin films using Monte Carlo simulation of fluorescence decay. *Chemical Physics Letters.* 2001;345(1–2):141–150. DOI: 10.1016/S0009-2614(01)00867-3.

34. Yatskou MM, Donker H, Novikov EG, Koehorst RBM, van Hoek A, Apanasovich VV, et al. Nonisotropic excitation energy transport in organized molecular systems: Monte Carlo simulation-based analysis of fluorescence and fluorescence anisotropy decay. *The Journal of Physical Chemistry A.* 2001;105(41):9498–9508. DOI: 10.1021/jp0044227.

35. Giganti A, Friederich E. The actin cytoskeleton as a therapeutic target: state of the art and future directions. In: Meijer L, Jézéquel A, Roberge M, editors. *Cell cycle regulators as therapeutic targets.* Roscoff: Life in Progress Editions; 2003. p. 511–525 (Progress in cell cycle research; volume 5). PMID: 14593746.

36. Pollard TD. Actin and actin-binding proteins. *Cold Spring Harbor Perspectives in Biology.* 2016;8(8):a018226. DOI: 10.1101/cshperspect.a018226.

37. Kitamura A, Kinjo M. State-of-the-art fluorescence fluctuation-based spectroscopic techniques for the study of protein aggregation. *International Journal of Molecular Sciences.* 2018;19(4):964. DOI: 10.3390/ijms19040964.

38. Halavatyi AA, Nazarov PV, Medves S, van Troys M, Ampe C, Yatskou M, et al. An integrative simulation model linking major biochemical reactions of actin-polymerization to structural properties of actin filaments. *Biophysical Chemistry.* 2009;140(1–3):24–34. DOI: 10.1016/j.bpc.2008.11.006.

39. Al Tanoury Z, Schaffner-Reckinger E, Halavatyi A, Hoffmann C, Moes M, Hadzic E, et al. Quantitative kinetic study of the actin-bundling protein L-plastin and of its impact on actin turn-over. *PLOS One.* 2010;5(2):e9210. DOI: 10.1371/journal.pone.0009210.

40. Lisitsa EV, Yatskou MM, Apanasovich VV, Apanasovich TV, Shytsik MM. Simulation model for three-channel luminescent images of cancer cell populations. *Journal of Applied Spectroscopy.* 2015;81(6):996–1003. DOI: 10.1007/s10812-015-0041-z.

41. Lisitsa YV, Yatskou MM, Apanasovich VV, Apanasovich TV. Algorithm for automatic segmentation of nuclear boundaries in cancer cells in three-channel luminescent images. *Journal of Applied Spectroscopy.* 2015;82(4):634–643. DOI: 10.1007/s10812-015-0156-2.

42. Yatskou MM, Apanasovich VV. Computational platform FluorSimStudio for processing kinetic curves of fluorescence decay using simulation modeling and data mining algorithms. *Journal of Applied Spectroscopy.* 2021;88(3):571–579. DOI: 10.1007/s10812-021-01211-6.

43. Yatskov NN, Apanasovich VV, Koehorst RBM, van Hoek A, Schaafsma TJ. Electronic spectra and fluorescence polarization kinetics of thin Zn-porphyrin films. *Journal of Applied Spectroscopy.* 2003;70(3):372–377. DOI: 10.1023/A:1025177320845.

44. Yatskou MM, Meyer M, Huber S, Pfenniger M, Calzaferri G. Electronic excitation energy migration in a photonic dye-zeolite antenna. *ChemPhysChem.* 2003;4(6):567–587. DOI: 10.1002/cphc.200300567.

45. Yatskou MM, Skakun VV, Nederveen-Schippers L, Kortholt A, Apanasovich VV. Complex analysis of fluorescence intensity fluctuations of molecular compounds. *Journal of Applied Spectroscopy.* 2020;87(4):685–692. DOI: 10.1007/s10812-020-01055-6.

46. Skakun VV, Yatskou MM, Nederveen-Schippers L, Kortholt A, Apanasovich VV. Component analysis of photon counting histograms in fluorescence fluctuation spectroscopy experiments. *Journal of Applied Spectroscopy.* 2022;89(5):930–939. DOI: 10.1007/s10812-022-01450-1.

Received 07.12.2023 / revised 07.12.2023 / accepted 15.12.2023.