A Digital Platform for Processing Fluorescence Spectroscopy Data Using Simulation Modelling and Machine Learning Algorithms

Mikalai Yatskou
Dept. of Systems Analysis and Computer Modelling
Belarusian State University
Minsk, Belarus
yatskou@bsu.by

Vladimir Apanasovich
Dept. of Systems Analysis and Computer Modelling
Belarusian State University
Minsk, Belarus
apanasovichv@gmail.com

Abstract. **A digital computational platform is proposed for processing fluorescence spectroscopy data, which implements complex analysis of experimental information based on the simulation modelling and machine learning algorithms. Data analysis includes partitioning biophysical data into clusters according to the degree of likeness in 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 medoids 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 feature spaces is done with the aim of user interaction. As a demonstrative example, the platform FluorSimStudio is implemented for processing time-resolved fluorescence measurements (https://dsa-cm.shinyapps.io/FluorSimStudio). The digital platform is an open system and allows addition of complex analysis models, taking into account the development of new modelling and analysis algorithms.**

Keywords: **fluorescence spectroscopy, simulation modelling, machine learning, digital platform**

I. INTRODUCTION

Experimental fluorescence spectroscopy methods are applied to study the optical properties of molecular compounds and are commonly used in the studies of artificial photonic materials, protein complexes, biopolymers, DNA sequencing, biological membranes, cell and tissues, medical diagnostics [1]. The considerable development of methods is driven due to the improvements of effective molecular fluorophores, including genetically expressed proteins (for example, GFP), semiconductor nanoparticles and quantum dots, optical systems for laser excitation and registration of radiation, allowing high-precision measurements, computer technologies for data storage and processing [2]. Novel experimental high-throughput techniques, integrating pulsed, phase and modulation methods for recording fluorescence decay times, form the basis of modern fluorescence microscopy and allow obtaining big data, characterized by high spectral, time and spatial resolution [3]. The main fluorescence spectroscopy and microscopy techniques for studying complex molecular systems in "cuvettes" and living cells are fluorescence-lifetime imaging microscopy (FLIM), fluorescence recovery after photobleaching (FRAP) and its derivatives – fluorescence loss in photobleaching (FLIP) and fluorescence localization after photobleaching (FLAP), fluorescence fluctuation spectroscopy (FFS, combining fluorescence correlation spectroscopy (FCS), fluorescence cross-correlation spectroscopy (FCCS), photon counting histogram (PCH) and fluorescence intensity distribution analysis (FIDA)), fluorescence sensing (FS) [4].

The existing data analysis approaches to processing fluorescence spectroscopy data can be divided into classical and modern, based on machine learning, 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 [5]. 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 [5, 6]. The main disadvantages of existing data processing methods are limited or poor efficiency, that is due to the use of nonphysical analytical models (multi-exponential or polynomial decompositions), poor accuracy in parameter estimating when analyzing 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

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

video cards or multiprocessor nodes to accelerate neural network computing), and finally the lack of specialized software for automated data processing. Therefore, the primary task is to develop an integrated data analysis approach and computational platform 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.

A computational approach for processing large sets of time-resolved fluorescence data using simulation modelling and data mining algorithms was developed [7, 8]. By this methodology it is possible to increase the accuracy of the estimated parameters of biophysical and optical processes occurring in the studied molecular systems. Specialized and general-purpose software tools and products, both commercial and freely available, have been developed for statistical processing, analysis and simulation of fluorescence spectroscopy data. However, there are no unified integrated software tools for processing large datasets using simulation modelling and machine learning methods. The development of a digital software platform for simulation and machine learning analysis of fluorescence data in various biophysical systems under experimental studies is an critically important and urgent task.

In this paper, we propose the conception of a digital software platform for the simulation modelling and machine learning analysis of optical processes in molecular systems studied by the fluorescence spectroscopy methods. As a demonstrative example, developed integrated methodology is implemented into the computational platform FluorSimStudio for processing fluorescence kinetic curves obtained through FLIM experiments.

II. Methodology

A. *Review of the Computational Tools for a Digital Platform.*

A digital computational platform in this case is an intellectual software resource or a programming environment designed to model and analyze large experimental fluorescence spectroscopy data studied in biophysical research. The platform includes a programming environment, integrated coding languages, software tools for automation, code debugging and creating an application interface, models of research objects, methods for analyzing and visualizing data, assessing the quality of analysis and the reliability of models. The choice of the optimal software platform primarily implies the choice of a programming environment and interface development tools for interacting with the user.

Various computing platforms and programming technologies are used to implement the software. In most publications on benchmarking open access packages, there is no clear leader in machine learning and data mining. Currently, a large number of software tools are actively used, including WEKA, Tanagra, Rapid Miner, KNIME, Orange, Java, Python and R projects, as well as platforms implemented using high-performance programming languages C++ and Scala. The advantages of these software resource are computational performance, a wide range of libraries for statistical analysis, cross-platform integrity, the ability to develop user interfaces, parallel computing, work directly with existing databases and data warehouses. The main disadvantages include the lack of versatility, significant requirements for computing resources, and the limitation of the integration of the above fascinating properties in a single format. The most promising projects for organizing the digital environment are Scala-, Python- and R-platforms. A platform based on the Scala language (for example, Apache Hadoop) is designed to analyze big data in production projects and is used to solve industrial programming problems. Python applications are aimed at solving general engineering and data analysis problems with an emphasis on neural network approaches and programming. R-projects are developed primarily with the aim of optimizing and validating applied statistical analysis, which includes approaches using classical and data mining methods. Let take a closer look at the R environment.

The main advantages of the statistical programming environment R are the presence of optimized structures for representing data objects, which greatly simplifies data processing, optimization of programming tools and implementation of computation algorithms (in the sense of minimizing the introduction of errors into the program code), the ability to use a huge set of processing algorithms, statistical and data mining, various computing resources of the scientific community [9]. The main drawback is the low computational performance in the basic version of the environment layout, which is especially critical when working with large datasets and developing simulation models. This limitation can be partially or completely eliminated by connecting program codes of high-performance programming languages Scala, Java, C++ (packages rscala, rjava, Rcpp, inline), parallel computing procedures (managed by packages parallel, Rmpi, snow, snowfall), additional packages for efficient processing big data (readr, LaF, data.table, ff, bigmemory) and the use of third-party software resources (Microsoft R Open and Intel Math Kernel Library libraries, H2O big data analysis platforms, Apache Hadoop and Spark systems, with using h20, Rhadoop and SparkR packages).

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

An important issue is the development of the interface of a software application. The most popular R-code integrating user interface development packages are gWidgets, rpanel, svDialogs, RGtk2, qtbase, tcltk. A new direction in the development of R-applications for the analysis of biophysical systems [10] is associated with the creation of "reactive" web interfaces using the Shiny package and the subsequent placement of the software implementation on the shinyapps.io resource provided by the open source software developers RStudio. The advantage of this approach is the ability to remotely work with a web application for a wide scientific audience of users online via the global Internet. To implement the software application, the R computing environment and the Shiny package were chosen to create a web interface for the developed application.

The computing platform is organized according to the example of open projects of network resources CRAN (https://cran.r-project.org), R-Forge (https://r-forge.r-project.org), Bioconductor (https://www.bioconductor.org), Github (https://github.com). It is a programming and simulation environment that contains updated and supplemented libraries of analytical and simulation models of optical processes in molecular systems, built-in tools for machine learning methods and assessment of the quality of analysis and modelling, provides the scientific community with opportunities to develop new algorithms and simulation models.

B. Conception of the Digital Platform.

The digital platform can integrate the research scheme for a certain biophysical process or molecular compound using a complex approach based on simulation modelling and machine learning methods [7]. A schematic diagram of the methodology for spectral or/and time-resolved fluorescence spectroscopy data analysis of the platform is shown in Fig. 1. Consider the main stages of data analysis.

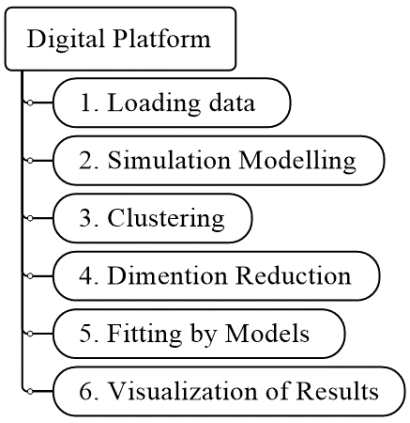

Fig. 1. Main stages of the fluorescence data analysis of a digital platform using simulation modelling and machine learning

The platform is designed to analyze experimental or simulated data. Data loading and graphical presentation is carried out in block 1. Visual assessment of two-dimensional and three-dimensional fluorescence datasets allows predetermining the choice of a mathematical model for describing the physical processes, making a supposition regarding the number of data clusters, and limiting the choice of measures for calculating the similarity of samples based on the noise level of the data.

Modelling and visualization of fluorescence data are carried out in block 2. Integrated models of optical processes are considered. Simulation modelling is carried out using Monte Carlo algorithms [11]. The input characteristics of the simulation are the type and parameters of the model, the number of samples and the number of simulations. 2D or 3D visualization are intended for expert analysis of modeled data, study of the behavior of models when changing their parameters, manual selection of the most optimal modelling parameters, such as the number of simulations and data points, as well as initial approximations of parameters for subsequent precise determination during fitting using mathematical models. New and improved models of optical-physical processes in molecular systems can be developed and integrated into the software environment.

In block 3, cluster analysis of fluorescence data is performed in the space of experimentally detected features. Clusters of data are identified according to some degree of similarity (Euclidean, Minkowski, Manhattan, maximum or Canberra distance). The number of clusters is determined intuitively, automatically from the hierarchy dendrogram of data constructed on the basis of the cluster binding measure (Ward, nearest neighbor, far neighbor, or middle bond), or on the basis of a statistical criterion [12, 13]. The median representatives of the clusters are calculated – medoids, samples or data objects having the smallest average distances to the rest of the objects of the corresponding clusters.

A data reduction is carried out in block 4. The consideration of a large group of uninformative experimentally detected features leads to difficulties in data analysis, namely, to their noise, an increase in the amount of data, and distortion of reliable information about clusters of similar samples. To improve the quality of data analysis, in particular, the visual assessment of data partitioning into clusters, it is expedient to carry out the stage of data analysis, which includes the transition to a low-dimensional space of new informative features, in which the fluorescence data form clusters. To perform this transformation, it is required to use data dimensionality reduction algorithms, among which

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

the method of principal component analysis is the most widely known [14]. Conversion of fluorescence data using principal component analysis is performed. The proportion of relative variation attributed to principal components is set, limiting the number of components. Principal components are selected that correspond to a given variation in the data (for example, 0.95 out of 1). A diagram of the proportions of variation of the first ten principal components is constructed, according to which the contribution to the total variance in the data is estimated. Clusters and their medoids are displayed in the scatter diagram of the first two principal components. Medoids are calculated in the space of initial features or in the space of the main components that explain a given value of variability. For example, if the data clusters are not separated, then it can be assumed that there is only one kind of fluorescent compounds. Otherwise, the presence of several forms of compounds (fluorophores) is allowed. For the convenience of visual control of cluster separability, histograms of frequencies are plotted on the axis of the first three principal components. Good separability of clusters is characterized by the presence of a multimodal form of histogram distributions.

In block 5, cluster medoids are analyzed to accurately determine the parameters of fluorescent compounds using an optimization algorithm and mathematical models. To approximate the fluorescence data, represented by the found medoids, analytical and simulation models for describing photophysical processes are used. Optimization methods are applied for the optimal selection of the parameters of mathematical models during the approximation of experimental data. In this work, the Nelder–Mead method [15] is chosen, which does not take into account the derivative of the objective function, which greatly simplifies the use of simulation models in the parameter estimation procedure. The best approximation is chosen according to a criterion (or a set of criteria) that determines the degree of deviation of the theoretical model from the experimental data. As a rule, such a criterion is presented analytically in the form of a function of experimental and theoretical data, the form of which is determined by the field of application, the direct modelling method and the conditions of the experiment. In our experiments, we consider the normalized chi-square criterion, diagrams of weighted residuals and their autocorrelation function [12].

The visualization of the results and the analysis of graphical images of the estimated data clusters are carried out with the aim of interpreting, explaining, improving the understanding of the research object and its behavior (block 6). Reduced data are plotted in the three principal component space, the original feature space, and the principal component coordinates that explain the given fraction of the variation in the data. The presentation of a diagram of three main components, interactive for user interaction, allows to visually assess the proximity of the found clusters and their shapes, the location of individual data points, the influence of experimental effects. Diagrams of a set of informative components enable to determine data clusters for a possible assessment of the parameters of models in the space of the main components. The latter helps to improve the accuracy of parameter estimation by reducing noise in the data due to the elimination of uninformative components describing the experimental noise. The procedure for estimating the parameters of models in the space of principal components can be additionally implemented in the platform. An interactive domain data cluster diagram let to qualitatively explore groups of processed data.

III. Results

For the practical implementation of the digital platform conception, integrating simulation modelling and machine learning algorithms, the computational platform FluorSimStudio is developed for processing fluorescence kinetic curves at FLIM experiments. It is launched on an R server hosted on a network resource, such as shinyapps.io. To implement simulation models, it is proposed to use the C++ programming language. The choice and development of algorithms for data analysis is carried out by direct programming or by connecting ready-made machine learning packages provided by the scientific community of developers through open projects CRAN, Bioconductor, Github. The user's work is carried out through a web application. In the structure of the computational approach, the platform integrates the implementation of simulation models, analysis algorithms, provides computational tools for applying the developed simulation models and methods to the analysis of datasets, instruments for assessing its quality, visualizing and interpreting data.

The programming implementation of the platform FluorSimStudio is organized using the Shiny R package and contains a set of functions that integrate the methodology for an integrated approach to data analysis. The web application is hosted at https://dsa-cm.shinyapps.io/FluorSimStudio. An example of the interface window is shown in Fig. 2. 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 (PCA), fitting medoids (data analysis), visualizing and interpreting the results, information about the authors of the development, and instructions for using the computational resource.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

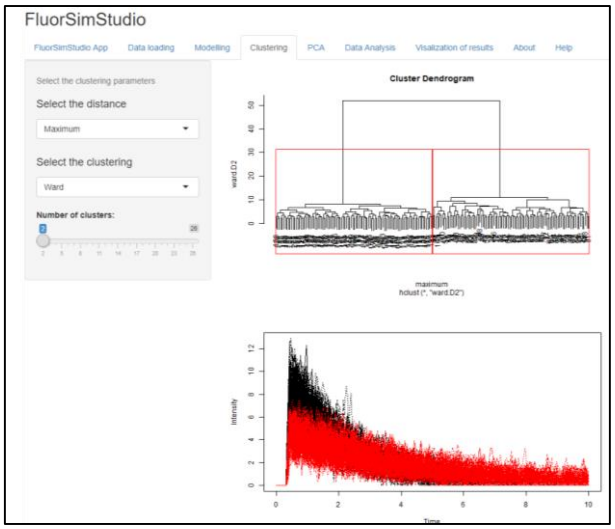

Fig. 2. FluorSimStudio web application interface window. Example of clustering fluorescence decay curves

The performance of the computational platform FluorSimStudio was tested by examples of the analysis of datasets representing systems of free fluorophores and in the presence of the Förster electronic excitation energy transfer process [1]. The obtained results are in good agreement with those previously published for analytical models of single- and stretch-exponential fluorescence decay laws [7]. Comprehensive analysis using simulation models and machine learning lets successfully to restore the parameters of optical processes from the experimental data.

IV. Conclusions

The conception of a digital platform for processing fluorescence spectroscopy data is developed, which is an implementation of an integrated approach for the complex machine learning analysis and modelling of optical processes in biophysical systems. Integrated data analysis pipeline comprises partitioning data into clusters, finding the cluster medoids, applying the data reduction method and visualizing the experimental data in a two-dimensional space, analyzing the medoids with analytical or simulation models. By this data analysis approach, it is possible to enhance the efficiency of the biophysical research. The digital platform is a programming environment designed to model and analyze large experimental fluorescence spectroscopy data. It includes a development framework, coding languages, tools for automation, code debugging and creating an application interface, models and methods for processing and visualizing data, assessing the quality of analysis. The R computing environment and the Shiny package are selected to create a web interface and online version for the developed software application. The C++ programming language is used for accelerating simulation modelling algorithms. The proposed methodology of the digital platform is realized in the computational platform FluorSimStudio, intended for processing fluorescence decay curves in molecular systems. FluorSimStudio 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 experimental systems. Computational efficiency of the digital platform can be increased by connecting software tools for high performance big data computing (for example, H2O, Apache Hadoop, Spark resources).

References

[1] J. R. Lakowicz, Principles of Fluorescence Spectroscopy, 3rd ed. Springer, New York, 2006.

[2] Z. Gryczynski, I. Gryczynski, Practical Fluorescence Spectroscopy. CRC Press, Boca Raton, 2019.

[3] G. Cox, Ed., Fundamentals of Fluorescence Imaging. Jenny Stanford Publishing Pte. Ltd, Singapore, 2019.

[4] D. M. Jameson, Introduction to Fluorescence. CRC Press, Boca Raton, 2014.

[5] R. Datta, T. M. Heaster, J. T. Sharick, A. A. Gillette, M. C. Skala, “Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications”, J. Biomed. Opt., vol. 25(7):071203:1-43, May 2020.

[6] R. Datta, A. Gillette, M. Stefely, M. C. Skala, “Recent innovations in fluorescence lifetime imaging microscopy for biology and medicine”, J. Biomed. Opt., vol. 26(7):070603:1-11, July 2021.

[7] M. M. Yatskou, V. V. Skakun, V. V. Apanasovich, “Method for processing fluorescence decay kinetic curves using data mining algorithms”, J. Appl. Spectr., vol. 87(2), pp. 333–344, May 2020.

[8] M. M. Yatskou, V. V. Skakun, L. Nederveen-Schippers, A. Kortholt, V. V. Apanasovich, “Complex Analysis of Fluorescence Intensity Fluctuations of Molecular Compounds”, J. Appl. Spectr., vol.87(4), pp. 685–692, September 2020.

[9] R. Gentleman et al., "Bioconductor: open software development for computational biology and bioinformatics", Genome Biology, vol. 5(10):R80, September 2004.

[10] V. Yuan, D. Hui, Y. Yin, M. S. Peñaherrera, A. G. Beristain, W. P. Robinson, "Cell-specific characterization of the placental methylome", BMC Genomics, vol. 22(1):6, January 2021.

[11] M. M. Yatskou, Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Systems. Wageningen University, The Netherlands, 2001.

[12] M. M. Yatskou, Data Mining [in Russian]. BSU, Minsk, 2014.

[13] H. Shimodaira, "Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling", Annal. Statist., vol. 32(6), pp. 2616–2641, December 2004.

[14] T. Jolliffie, Principal Component Analysis. Springer, New York, 2002.

[15] J. A. Nelder, R. Mead, "A simplex method for function minimization", Comput. J., vol. 8(1), pp. 308–313, January 1965.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.