Complex Analysis of Fluorescence Intensity Fluctuations of Molecular Compounds

Yatskou, M. M.; Skakun, V. V.; Nederveen-Schippers, L.; Kortholt, A.; Apanasovich, V. V.

Published in:
Journal of applied spectroscopy

DOI:
10.1007/s10812-020-01055-6

Publication date:
2020

DOI 10.1007/s10812-020-01055-6
Journal of Applied Spectroscopy, Vol. 87, No. 4, September, 2020 (Russian Original Vol. 87, No. 4, July–August, 2020)

COMPLEX ANALYSIS OF FLUORESCENCE INTENSITY FLUCTUATIONS OF MOLECULAR COMPOUNDS

M. M. Yatskou,[a*] V. V. Skakun,[a] L. Nederveen-Schippers,[b] A. Kortholt,[b] and V. V. Apanasovich[c]

UDC 535.37:547.96

A method is proposed for the complex analysis of fluctuations in the fluorescence intensity of molecular compounds, which allows determining the structural composition of protein oligomers. The idea of the method is to analyze the photon counting histograms of experimental measurements using principal component analysis to assess the presence of oligomeric compounds, and to perform hierarchical cluster analysis, to determine the data classes corresponding to various molecular compounds, followed by selecting cluster medoids to determine the oligomeric composition of protein complexes. The efficiency of the analysis algorithms developed within the framework of the proposed method was confirmed on simulated and experimental photon counting histograms of the measured fluorescence intensity fluctuations of monomeric and dimeric forms of green-fluorescent protein (GFP).

Keywords: *fluorescence intensity fluctuation, photon counting histogram, molecular compounds, protein oligomers, data mining, principal component analysis, hierarchical cluster analysis, green-fluorescent protein (GFP).*

Introduction. Fluorescence fluctuation spectroscopy is widely used to study the diffusion of proteins and their interactions in living cells [1–3]. In the course of the experiment, the fluorescence of molecules bound or freely moving in a solution or a cell is recorded in a certain small volume (up to 10^{-18} m^3) formed by an extremely focused laser beam. Fluctuations in fluorescence intensity are primarily due to changes in the number and location of molecules in the recorded volume, as well as their interaction and the properties of the medium. The oligomeric composition of a protein compound can be determined by analyzing the amplitude of fluctuations in fluorescence intensity over time (methods for analyzing the distribution of fluorescence intensity — PCH (photon counting histogram) [4] and FIDA (fluorescence intensity distribution analysis) [5]). In the PCH and FIDA methods, a histogram of the number of photocounts (PC) is plotted at a given recording time interval to determine the concentration of a protein freely emitting or labeled with a luminescent dye. The recorded fluorescence intensity of the sample is directly proportional to the number of fluorescent molecules that form the studied molecular complex, which makes it possible to estimate the number of molecules inside the protein complex and the size of the complex [6, 7].

To analyze the distribution of the number of photocounts, various mathematical models [4–7] and optimization methods are usually used, among which the least squares method with Levenberg–Marquardt optimization [8] is used most often, which makes it possible to obtain information on the diffusion and structural properties of the studied protein compounds in the first approximation. However, the classical iterative algorithms for data analysis have a number of significant limitations. They do not allow one to accurately determine the number and type of molecular oligomers, perform a local rather than global search for model parameters, and require significant computational costs for data analysis. An alternative approach to solving this problem is the use of mining algorithms and large multidimensional data, the essence of which is the simultaneous global analysis of the entire data set as a whole [9–12].

In the present work, we propose a method for the complex analysis of fluorescence intensity fluctuations and the PCHs based on them using intelligent analysis algorithms in order to determine the oligomeric composition of molecular compounds.

*To whom correspondence should be addressed.

[a]Department of Systems Analysis and Computer Modelling, Belarusian State University, Minsk, 220030, Belarus; email: yatskou@bsu.by; [b]University of Groningen, 9747AG Groningen, The Netherlands; [c]Institute of IT & Business Administration, Minsk, 220004, Belarus. Translated from Zhurnal Prikladnoi Spektroskopii, Vol. 87, No. 4, pp. 628–636, July–August, 2020. Original article submitted February 4, 2020.

Methodology. The developed method is based on the hypothesis of the separability of a set of multidimensional experimental data in a certain information space into several populations representing various molecular oligomeric compounds [10]. A small measurement volume is considered, in which molecular compounds of the same type prevail in a series of short time intervals. A normal distribution of the measured attributes is assumed for molecular compounds of the same type in the allocated space. For example, protein monomers can form a cloud or spherical Gaussian cluster of data in a multidimensional space based on measurable attributes. If, however, protein oligomers are added to the monomeric forms of the protein, then the cloud is extended or divided into two parts along a certain line connecting the centers of the two populations. In the extreme case, two clouds or clusters of these monomers and oligomers are expected. Thus, if groups of data are divided into clusters in a multidimensional space of attributes, this confirms the presence of several forms of protein compounds. Tasks of this kind are solved using data mining algorithms such as data dimensionality reduction and cluster analysis [10, 13, 14]. Dimensionality reduction algorithms allow switching to a low-dimensional space without losing the essence of information [15, 16]. Cluster analysis algorithms make it possible to determine clusters of data specified in varying degrees of similarity, the number of which may be associated with aggregates of molecular compounds. Thus, applying principal components analysis (PCA) will make it possible to carry out such a rotation, as a result of which the axis of the first principal component coincides with the diagonal of the data cloud in multidimensional space [17]. Therefore, the relative fraction of the scatter attributable to the first principal component for two types of molecular compounds (an elongated ellipsoid or two spherical data clouds in a multidimensional space of attributes is expected) should differ significantly from that for a monomer solution (one spherical cloud). It should be noted that the scatter diagram of the first two principal components is informative in the sense of defining the data structure in two-dimensional space.

The idea behind the method of complex analysis is to calculate the PCH based on the recorded fluorescence intensities (it is possible to use other attributes, for example, the autocorrelation function or factorial cumulants of the distribution of the number of photocounts [18]), the use of the PCA to assess the presence of oligomeric compounds and hierarchical cluster analysis to determine groups of data, corresponding to various molecular compounds, followed by the isolation of cluster medoids, PCHs having the smallest average distances to the remaining objects of the corresponding clusters, to assess the parameters of the oligomeric composition of protein complexes. Comprehensive analysis requires the availability of experimental data for the reference (monomers) and tested (oligomeric forms) samples. The block diagram of the developed method is shown in Fig. 1. Consider the main stages of the method.

Calculation of the PCH. We calculate N of the PCH based on the registered sets of fluorescence intensities S_i, $i = 1, 2, ..., N$, and form objects $n_1, n_2, ..., n_N$, characterized by attributes $X_1, X_2, ..., X_K$, — histogram channels representing the frequencies of occurrence f_j of the number of photons $l = (j - 1)$, $j = 1, 2, ..., K$, during a certain (short) time interval Δt. As a standard or reference sample, we use the experimental data of the monomer solution, and as a test sample — data for the oligomeric forms of the protein.

Data Dimensionality Reduction. The PCA method is applied to datasets of reference and test samples. In the PCA, such a linear transformation is defined, as a result of which the initial data $X_1, X_2, ..., X_K$ are expressed by a set of principal components $Z_1, Z_2, ..., Z_K$, where the first M principal components ($M << K$) provide the required fraction γ of the variance of groups of attributes. In expanded form, the principal component Z_j is expressed through the attribute vectors $X_1, X_2, ..., X_K$:

$$Z_j = a_{1j} X_1 + a_{2j} X_2 + ... + a_{Kj} X_K , \tag{1}$$

where a_{ij} are the loading parameters of the principal components. The relative proportion of the scatter (%) attributable to the principal component Z_j is:

$$\alpha_j = 100 \frac{D(Z_j)}{D(Z_1) + D(Z_2) + ... + D(Z_K)} , \tag{2}$$

where $D(Z_j)$ is the variance of the component Z_j. If the relative proportions of the scatter in the reference and the tested samples, which fall on the first principal component Z_1, are the same, then to assume that there are no oligomers means to stop the algorithm. Otherwise, permit the presence of oligomers and continue the algorithm.

Hierarchical Cluster Analysis of the Reference Sample (HCARS). A hierarchical cluster analysis of the histograms of the reference sample $n_1^R, n_2^R, ..., n_N^R$ is performed in the space of initial attributes. In this case, it is necessary to specify a method for comparing objects to each other (or a measure of similarity, for example, Euclidean, Minkowski, correlation distance). In the developed method to eliminate inter-experimental inhomogeneities associated with separate measurements

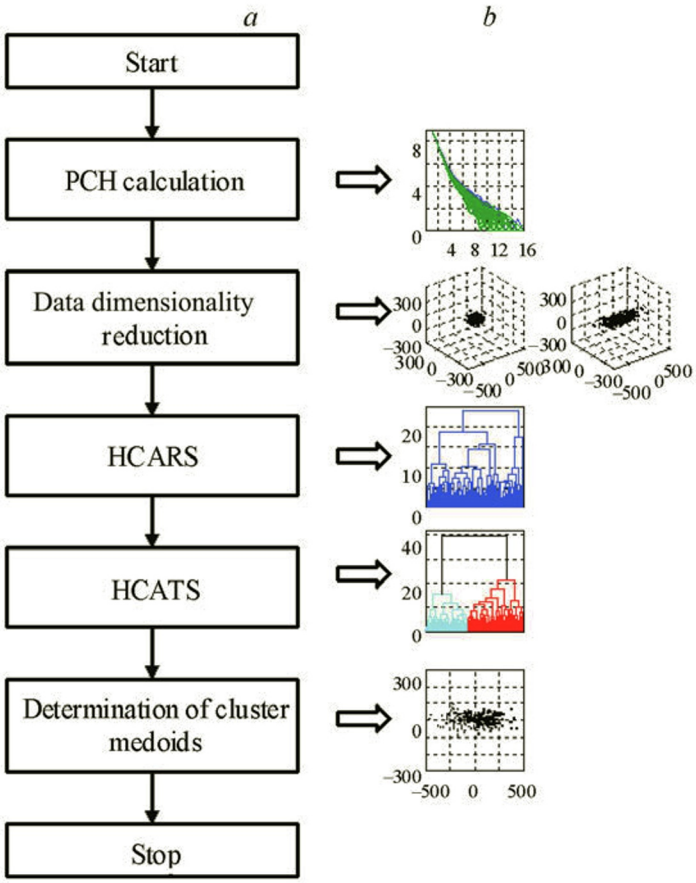

Fig. 1. Block diagram of the method (a) and diagram of the results of its main stages (b) for studying fluctuations of the fluorescence intensity of molecular compounds using data mining algorithms.

of the reference and test samples, we propose to use the standardized Euclidean distance (invariant to inhomogeneity in the data) [10]:

$$d_e(n_i, n_j) = \sqrt{\sum_{l=1}^{K} \frac{(x_{il} - x_{jl})^2}{\sigma_l^2}}, \quad (3)$$

where x_{il} and x_{jl} are coordinates of objects n_i and n_j; σ_l^2 is the variance of the attribute X_l. We determine the maximum connection distance (or threshold) d_1 on the dendrogram, at which the data are combined into one cluster. The maximum connection distance d_1 is used as a threshold for finding the number of oligomer clusters on the dendrogram for the test data.

Hierarchical Cluster Analysis of the Test Sample (HCATS). A hierarchical cluster analysis of the histograms of the tested sample n_1^T, n_2^T, …, n_N^T is performed in the space of initial attributes. Using the threshold d_1 found in the previous step of the algorithm, we select data clusters on the dendrogram. Assume that one cluster belongs to monomers, and the other(s) — to oligomeric forms.

Determination of Cluster Medoids. Clusters of monomers and oligomers are displayed on the scatter diagram of the first two principal components. Datasets are formed by calculating medoids in each cluster to accurately determine the parameters of molecular compounds using PCH and FIDA methods.

Materials and Methods. Consider simulated and experimental data. The simulated data make it possible to qualitatively assess the performance of the method and explore the limits of application. The experimental data are used to confirm the fundamental possibility of applying the developed approach to solving real problems of experimental research.

A simulation model of the photocount flow with a given distribution of the number of photocounts is presented in [19]. The number of photons emitted by the molecule during the observation time T is approximated by the Poisson distribution with the intensity

$$\lambda_f = \langle q \rangle TB(r) , \quad (4)$$

where $\langle q \rangle$ is the brightness, or the average number of photons emitted by one molecule per unit of time; $B(r)$ is the exposure profile function; $r(x, y, z)$ is the radius vector of the molecule. A three-dimensional Gaussian distribution is used as a function of the exposure profile $B(r)$. The number of molecules in solution in a certain volume obeys the Poisson distribution with the parameter

$$\lambda_m = \langle N_m \rangle V_0 , \quad (5)$$

where $\langle N_m \rangle$ is the average number of molecules of the test sample per unit volume; V_0 is the exposure volume. For each molecule, the coordinates of the location in the volume V_0 (according to the uniform distribution law) and the number of emitted photons (according to the Poisson distribution with the intensity λ_f) are generated. If a mixture of molecules of different types is simulated, then it is necessary to perform photon generation cycles for each type of molecule. The generation cycle is repeated iteratively until the accumulation of the number of photons, at which a PCH with a given signal-to-noise ratio is formed. To take into account the effect of scattering of data or "blurring" of PCH clusters caused by the influence of various distortions, such as the presence of unremovable impurities that quench or stimulate fluorescence of molecules, high background noise, flare and degradation of dyes, we use modeling of model parameters that have a normal distribution with a given mathematical expectation and standard deviation σ. Variation of σ makes it possible to control the scatter of data or the blur of clusters of PCH curves in a multidimensional space of time samples.

The simulated data is an example of an idealized system of two types of molecules: a monomer (M) and a dimer (D) of a certain protein (for example, GFP in solution), separately generated PCHs of which are characterized by the average number of molecules in the recording volume and their average brightness $\langle N^{\mathrm{M}} \rangle = 2$, $\langle q^{\mathrm{M}} \rangle = 5 \cdot 10^4$ and $\langle N^{\mathrm{D}} \rangle = 1$, $\langle q^{\mathrm{D}} \rangle = 10^5$. Observation interval is $T = 5 \cdot 10^{-5}$ s. Modeling was carried out with σ = 0.02 and 0.2 of the absolute values of the parameters $\langle N^{\mathrm{M}} \rangle$, $\langle q^{\mathrm{M}} \rangle$, $\langle N^{\mathrm{D}} \rangle$, and $\langle q^{\mathrm{D}} \rangle$.

Experimental data — well-known monomeric and dimeric forms of the green fluorescent protein GFP S65T [20] — were provided by the Cell Biochemistry Laboratory of the University of Groningen (Netherlands). Reference samples: GFP protein in buffered lysis solution (50 mM Tris, 50 mM NaCl, 5 mM DTT, 5 mM $MgCl_2$, 1% PI mix, 1% Triton X-100); separate measurements of the monomer (mGFP) and the stable dimer (diGFP, synthesized by liganding the pDM313 vector into pDM334 at the SpeI/XBaI binding sites) of GFP protein in lysates of dictyostelium cells. A test sample is a mixture of equal proportions of low concentrations ($\langle N_m \rangle < 1$) of mGFP and diGFP proteins in dictyostelium cell lysate. The measurements of the first sample were performed using a Leika TCS fluorescence confocal inverted microscope equipped with a lens immersed in oil (100×, 1.4NA) and a PicoHarp 300 (PicoQuant) photocount counting and recording system. The second and third samples were examined using a scanning inverted confocal microscope LSM 710 (Carl Zeiss) equipped with a lens immersed in water (100×, 1.2NA) and a Confocor3 measurement system (Carl Zeiss). The fluorescence of the samples was excited at λ = 488 nm and recorded in the λ = 505–610 nm range.

The simulated data make it possible to investigate the applicability of the developed method in the case of different separability of data clusters (varied by the parameter σ) corresponding to protein compounds. The data representing the GFP protein in the buffer solution and the cell lysate are experimentally confirmed and make it possible to check the efficiency of the method using examples of real model data. A mixture of monomeric and dimeric forms of the GFP protein is an example of a dataset specifically containing various forms of protein aggregation. Assuming that molecules of the same type were predominantly found in the observation volume, the PCHs of the experimental samples were constructed over a time interval of $5 \cdot 10^{-2}$ s or less in one measurement of fluorescence intensity fluctuations with a duration of 120 s.

The algorithms were implemented in the Matlab mathematical programming environment using the pdist, linkage, cluster, and eig functions, which integrate algorithms for hierarchical cluster analysis and PCA [21]. The hierarchical method of cluster analysis was used, and the most common method for calculating the distance (standardized Euclidean) and the measure of cluster similarity (Ward) were investigated [13]. The data centering procedure is applied in the PCA. To assess the error ε of restoring the PCHs of various types of molecules, the ratio of incorrectly determined PCHs to the total number of PCHs (in %) was considered.

Results and Discussion. The results of the analysis of the simulated datasets using the algorithms of the integrated approach are shown in Fig. 2 and in Table 1. The analysis of the simulated data was carried out separately for monomers

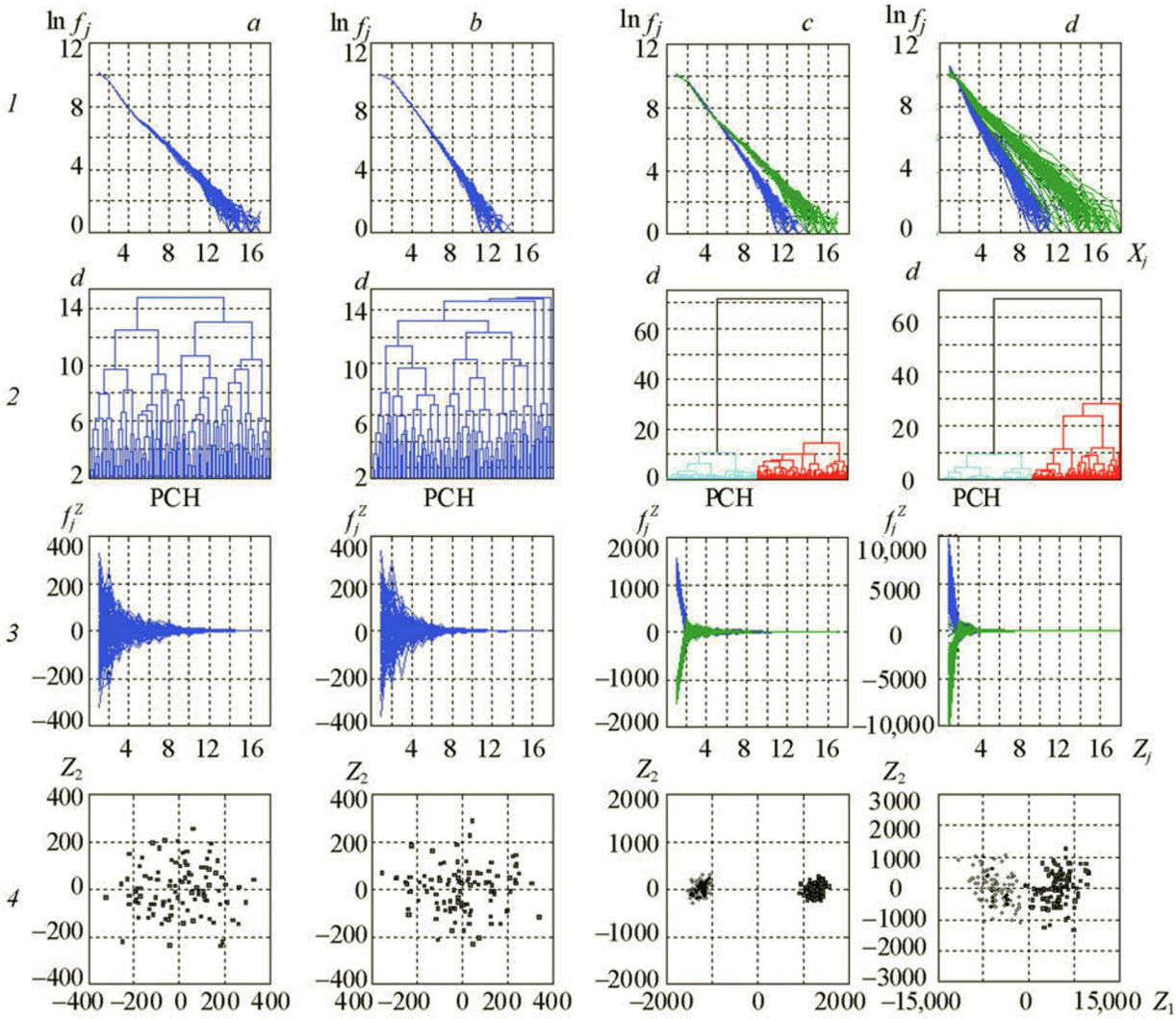

Fig. 2. The results of the analysis of the simulated data using the developed method, based on the algorithms of the principal component method (data centering is performed) and hierarchical cluster analysis (the standardized Euclidean measure of similarity of objects and the Ward connection distance for combining clusters are implemented); modeling parameters: $\langle N^{\mathrm{M}} \rangle = 2$, $\langle q^{\mathrm{M}} \rangle = 5 \cdot 10^4$ and $\langle N^{\mathrm{D}} \rangle = 1$, $\langle q^{\mathrm{D}} \rangle = 10^5$; a) monomers, $\sigma = 0.02$; b) dimers, $\sigma = 0.02$; c and d) combined sets of monomers and dimers with $\sigma = 0.02$ and 0.2; 1) photon counting histograms on a logarithmic scale in the space of the initial attributes $X_1, X_2, \ldots, X_K$; 2) dendrograms of photon counting histograms PCHs, d is the measure of cluster similarity; 3) photon counting histograms in the space of principal components $Z_1, Z_2, \ldots, Z_K, f_j^{\mathrm{Z}}$ — linearly transformed frequencies of occurrence of the number of photons in the coordinates of principal components; 4) histograms of photon counts in space of the first two principal components; the dimensionality of the axes of the principal components is represented by the linearly transformed frequencies of occurrence of the number of photons in the coordinates of components 1 and 2; shades of gray indicate monomeric and dimeric forms of proteins.

TABLE 1. Relative Proportion of Scatter (in %) for the First 10 Principal Components Obtained During Analysis of Simulated (SD) and Experimental Datasets Using Principal Component Analysis

Components	1	2	3	4	5	6	7	8	9	10
SD, monomers	**54.564**	29.263	8.134	3.079	2.318	1.120	0.701	0.471	0.166	0.094
SD, dimers	**58.775**	25.822	9.317	3.410	1.611	0.704	0.195	0.100	0.045	0.014
SD 1*	**98.768**	0.823	0.206	0.104	0.048	0.027	0.012	0.007	0.003	0.001
SD 2**	**98.998**	0.812	0.160	0.017	0.008	0.003	0.002	0.001	0.000	0.002
GFP	**50.502**	16.554	12.656	9.545	6.331	2.674	1.137	0.343	0.150	0.077
mGFP/diGFP	**99.869**	0.041	0.025	0.018	0.014	0.012	0.007	0.005	0.004	0.003
mGFP/diGFP mixture	**93.592**	4.161	1.360	0.470	0.175	0.104	0.055	0.028	0.023	0.011

*Monomers/dimers, $\sigma = 0.02$;

**Monomers/dimers, $\sigma = 0.2$.

and dimers (Fig. 2a and 2b). The relative proportion of the scatter α_1 for the first principal component is 54.6 and 58.8% for monomers and dimers, and the data clouds in the space of the principal components have a spherical Gaussian shape. The threshold value of the similarity measure, at which molecules form a single cluster $d_1 = 15$, is a criterion for determining clusters of different molecular shapes. The connection distance of the resulting clusters into one is <2, which indicates a significant similarity of the combined clusters.

The application of the algorithms of the developed method to the analysis of the combined set of simulated data makes it possible to accurately determine the samples of monomeric and dimeric forms of proteins (error $\varepsilon = 0$), which is confirmed by the high relative fraction of the scatter falling on the first principal component, α_1 >98% (for monomers 54.6%), clear separability of data into two clusters in the space of the principal components Z_1 and Z_2 (Fig. 2c), long connection distances of the resulting clusters into one (>50), which confirms the importance of the difference between clusters. It should be noted that the method successfully works under the conditions of the considered example of blurring and partial overlapping of data clusters ($\sigma = 0.2$, $\varepsilon = 1.5\%$; Fig. 2d), which is typical for molecular systems such as a mixture of GFP monomers and dimers in a cell lysate. Samples of monomeric and dimeric forms of proteins were determined: the relative proportion of scatter α_1 = 99%, the data form two clusters in the space of the principal components Z_1 and Z_2 (Fig. 2d), the line length of the unification of the resulting clusters into one is >30.

In the course of the study, together with the standardized Euclidean distance, three additional measures for calculating the similarity between objects, invariant to data heterogeneity, such as Mahalanobis, correlation and Spearman were considered [9, 13, 14]. The best results were obtained for the distances of the standardized Euclidean distance and Mahalanobis. However, the Mahalanobis measure requires the computation of the covariance matrix of the input data, which can be costly in the case of analyzing large datasets ($N \to \infty$, $K \to \infty$).

The results of the analysis of experimental datasets using the algorithms of the integrated approach are shown in Fig. 3 and in Table 1. Study of the data for the GFP protein in a buffer solution allows one to determine the threshold value of the similarity measure ($d_1 = 23$), at which the monomers form a single cluster, for use in the subsequent analysis of protein compounds (Fig. 3a). The connection distance of the resulting clusters into one (<5), the spherical shape of the data cloud in the space of the first two principal components (Fig. 3a) and a low relative proportion of the scatter $\alpha_1 = 50.5\%$ (Table 1), which falls on the first principal component, qualitatively confirm the fundamental principle of the working hypothesis proposed in the implemented method. As a result of the analysis of the combined experimental data of mGFP and diGFP proteins in cell lysates, the presence of two forms of proteins corresponding to monomeric and dimeric forms (Fig. 3b) was confirmed: $\alpha_1 = 99.9\%$, the data form two clusters in the space of the principal components, the connection distance of the resulting clusters into one >40. Analysis of the experimental data of a mixture of mGFP and diGFP proteins in the cell lysate revealed the presence of two forms of protein oligomers. The relative proportion of the scatter α_1, which falls on the first principal component of the tested data, at 93.6% significantly exceeds the value of 50.5% obtained for monomeric forms of the GFP protein in a buffer solution. The connection distance at which the final cluster is formed is 40

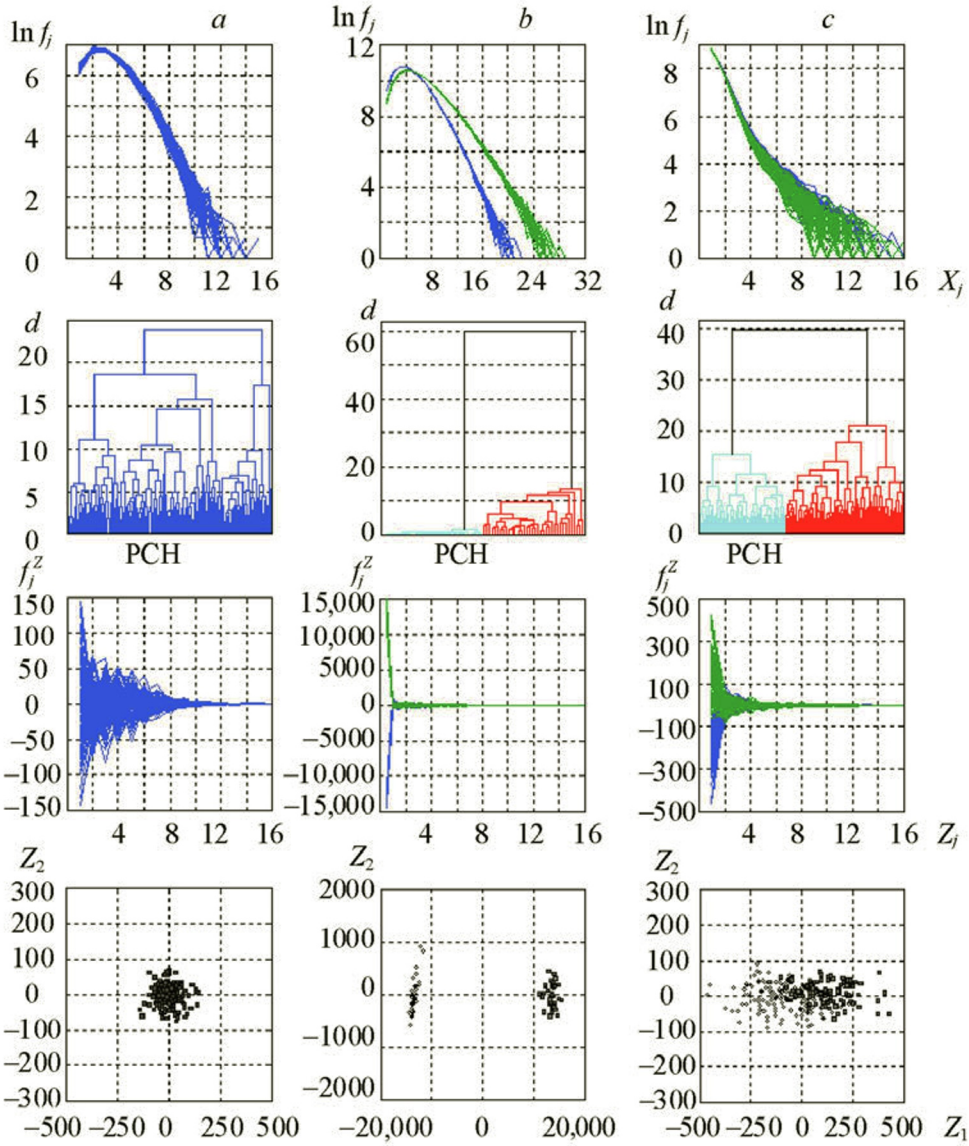

Fig. 3. Results of the analysis of experimental data sets using the developed method, based on the algorithms of the principal component analysis (data centering was performed) and hierarchical cluster analysis (the Euclidean measure of similarity of objects and the Ward connection distance for combining clusters were implemented): a) GFP protein in a buffer solution, b) mGFP and diGFP proteins in cell lysates, c) mixture of mGFP and diGFP proteins in cell lysates; designations as in Fig. 2.

(Fig. 3c), the data form two clusters in the space of the principal components, at 18 the connection distance of the resulting clusters into one significantly exceeds the value of 5 for GFP monomers. The value ≥23 should be taken as the threshold value for determining the number of nonmonomeric form clusters. At a connection distance of 23, two clusters formed by the majority of mGFP or diGFP molecules can be distinguished on the dendrogram of the tested data (Fig. 3c). Further evaluation of the parameters of protein complexes can be carried out in the course of analysis of medoids of the obtained PCH clusters using classical algorithms for analyzing fluorescence spectroscopy data [5, 6]. Note that the monomers of the GFP protein form a spherical cluster of data in the space of the first two principal components (Fig. 3a), while an elongated ellipsoidal cloud is observed for a mixture of mGFP or diGFP, formed by clusters of monomers and dimers of compounds (Fig. 3c).

Conclusions. A method for the complex analysis of fluctuations of the fluorescence intensity of molecular compounds is proposed, which makes it possible to determine the structural composition of protein oligomers and complements the classical methods of PCH and FIDA analysis. The efficiency of the algorithms developed within the framework of the proposed method was confirmed during the analysis of simulated and experimental data representing the fluorescence of monomeric and dimeric forms of the GFP protein. The developed method has the following advantages over the classical method for analyzing data from fluorescence fluctuation spectroscopy: it improves the accuracy of data analysis, since it uses the entire data set, rather than individual histograms; provides computational performance due to the high speed of execution of procedures of the method of principal components and cluster analysis in comparison with a separate analysis of the full set of histograms; provides the ability to visualize data in the space of the first two principal components, which is much more informative than a diagram of a complete set of initial histograms.

REFERENCES

1. E. L. Elson and D. Magde, *Biopolymers*, **13**, No. 1, 1–27 (1974).
2. A. Kitamura and M. Kinjo, *Int. J. Mol. Sci.*, **19**, No. 4, pii: E964, 1–18 (2018).
3. S. Veerapathiran and T. Wohland, *J. Biosci.*, **43**, No. 3, 541–553 (2018).
4. Y. Chen, J. D. Müller, P. T. So, and E. Gratton, *Biophys. J.*, **77**, 553–567 (1999).
5. P. Kask, K. Palo, D. Ullmann, and K. Gall, *Proc. Natl. Acad. Sci. USA*, **96**, No. 24, 13756–13761 (1999).
6. Y. Chen, L. N. Wei, and J. D. Müller, *Proc. Natl. Acad. Sci. USA*, **100**, No. 26, 15492–15497 (2003).
7. V. V. Skakun and V. V. Apanasovich, *Vestnik BSU. Ser. 1, Physics. Mathematics. Informatics*, **1**, 52–59 (2016).
8. D. Marquardt, *SIAM J. Appl. Math.*, **11**, No. 2, 431–441 (1963).
9. N. N. Yatskov, *Data Mining* [in Russian], BSU, Minsk (2014).
10. N. N. Yatskov, V. V. Skakun, and V. V. Apanasovich, *Applied Problems of Optics, Informatics, Radiophysics and Condensed Matter Physics* [in Russian], NII PFP BSU, Minsk (2019), pp. 122–124.
11. M. Bramer, *Principles of Data Mining*, Springer, London (2013).
12. C. C. Aggarwal, *Data Mining: The Textbook*, eBook, Springer (2015).
13. I. D. Mandel', *Cluster Analysis* [in Russian], Finansy and Statistika, Moscow (1988).
14. M. B. Lagutin, *Visual Mathematical Statistics* [in Russian], BINOM, Laboratoriya Znanij, Moscow (2007).
15. P. V. Nazarov, A. K. Wienecke-Baldacchino, A. Zinovyev, U. Czerwińska, A. Muller, D. Nashan, G. Dittmar, F. Azuaje, and S. Kreis, *BMC Med. Genom.*, **12**, No. 1, 132(1–17) (2019).
16. N. Sompairac, P. V. Nazarov, U. Czerwinska, L. Cantini, A. Biton, A. Molkenov, Z. Zhumadilov, E. Barillot, F. Radvanyi, A. Gorban, U. Kairov, and A. Zinovyev, *Int. J. Mol. Sci.*, **20**, No. 18, E4414 (1–27) (2019).
17. I. T. Jolliffie, *Principal Component Analysis*, Springer, New York (2002).
18. V. V. Skakun, E. G. Novikov, T. V. Apanasovich, and V. V. Apanasovich, *Methods Appl. Fluores.*, **3**, No. 4, 1–12 (2015).
19. I. P. Shingaryov, V. V. Skakun, and V. V. Apanasovich, *Methods Mol. Biol.*, **1076**, 743–755 (2014).
20. A. Kortholt, J. S. King, I. Keizer-Gunnink, A. J. Harwood, and P. J. M. Van Haastert, *Mol. Biol. Cell*, **18**, No. 12, 4772–4779 (2007).
21. N. N. Yatskov and E. V. Lisitsa, *Data Mining: Guidelines for Laboratory Work* [in Russian], BSU, Minsk (2019).