NEURAL NETWORK BASED ALGORITHM OF PRELIMINARY DATA ANALYSIS: APPLICATION TO FLUORESCENCE AND ESR SPECTROSCOPY

Nazarov P. V., Kavalenka A. A., Makarava K. U., Lutkovski V. M. and Apanasovich V. V.

Department of System Analysis, Belarusian State University, Skoryna Ave 4, 220050, Minsk, Belarus. e-mail: *nazarov@bsu.by*

Abstract. The fluorescence and electron spin resonance (ESR) spectroscopy are very important experimental tools for studying complex biomolecular objects and systems. The analysis of spectroscopic experimental data is often conducted by means of fitting using an analytical or a simulation models. For the successful performance of the fitting operation, an adequate model should be selected and good initial estimations of its parameters should be made. We propose to use artificial neural networks (ANNs) to recognise the appropriate model and to produce initial estimations of model's parameters before fitting.

1. Introduction

The fluorescence and electron spin resonance (ESR) spectroscopy are very important experimental tools for studying the complex biomolecular objects and systems, including lipids, membranes, proteins, DNA, etc. These methods provide detailed information about structural and dynamic properties of these systems [2, 3].

The data analysis of these spectroscopic techniques is rather complex because of several reasons: there could be a number of unknown parameters; almost all dependencies between them and the processes occur are non-linear; and experimental data are distorted by noises and inaccuracies of a registration system. These facts impel to analyse experimental data *via* the multi-parametric optimisation approach (fitting). General scheme of the proposed method is the following: the model that describes studied processes is selected from all possible models, initial estimations for the model parameters are made, and the optimization algorithm is starting to modify these parameters trying to achieve a coincidence between experimental and calculated data.

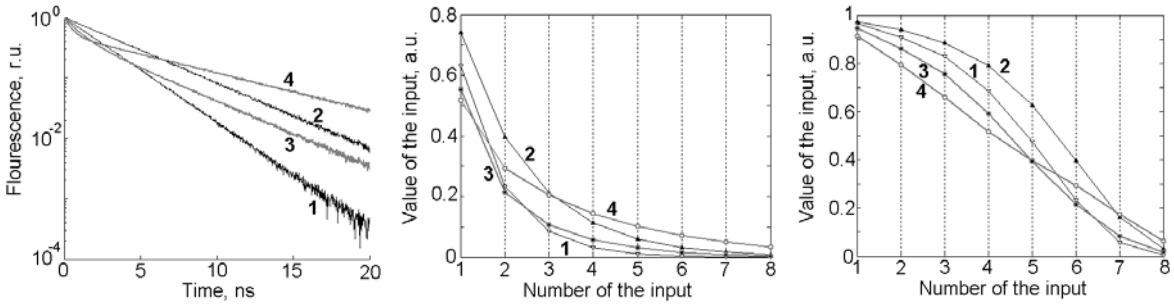
For the successful performance of the fitting procedure, the selected model must be an adequate and the initial estimation for its parameters should be sufficiently good. Consequently, the tasks of model recognition and initial estimations arise. The first task can be accomplished using *a priori* knowledge about the system. Unfortunately, it is not always possible, because this information may be the object of the research itself. To perform the second task, specific algorithms of data analysis can be implemented. However, these algorithms are strictly specialised, and cannot be applied in the general case. For example, the Laplace transform allows to analyse multi-exponential fluorescence decay model but it cannot be used for stretched exponential model.

Therefore, in this paper we propose to use artificial neural networks (ANNs) [1] to solve the tasks of model selection and initial parameter estimation. Neural networks are widely used in a variety of disciplines, including the application of such techniques to the data acquisition and triggering of high energy physics detectors. Their robustness provides successful data analysis in the presence of statistical fluctuations and noise.

2. General approach to ANN analysis

Several procedures have to be performed for the successful application of the neural network analysis. First, the dimensionality of the experimental data should be lowered by the mean of any approximation algorithm, for example – quantisation with the constant or logarithmic scale. From the variety of algorithms, the one, which saves most relevant information, should be selected. In addition, inputs of ANN should be scaled into similar ranges around the interval [0, 1].

Let us consider the pre-processing stage in the case of fluorescence decay analysis. The initial fluorescence decays are shown in the fig. 1*a*. Each experimental curve contains 1024 points – that is too much for an ANN. For the illustration, let the maximal number of ANN inputs be eight. The signals may be quantised into segments with constant (fig. 1*b*) or logarithmic (fig. 1*c*) length. In the second case, the signal was split into eight parts with the lengths of 8, 16, 32, 64, 128, 256, 512 points and the average value was taken for each part.



a) *b*) *c*)

Fig. 1. The examples of mono-exponential (curves 1, 2) and two-exponential (curves 3, 4) fluorescence decays (a) pre-processed via constant (b) and logarithmic (c) quantisation.

3. Data analysis in time-resolved fluorescence spectroscopy

Time-resolved fluorescence spectroscopy. Fluorescence spectroscopy is based on the emission of light from an electronically excited molecule [3]. After absorption of a photon the molecule turns into an excited electronic state. Excited molecule can return to its original energy state in several ways, including conversion of the energy into the heat (internal conversion), de-excitation by collision with another molecule (quenching), decomposition (photodissociation), or emission of light (fluorescence or phosphorescence). Fluorescence and phosphorescence differ in that fluorescence proceeds by emission from a singlet excited state to a singlet ground state (lifetime ~ nanoseconds) whereas phosphorescence proceeds by emission from a triplet state to a singlet ground state (lifetime ~ microseconds).

The most sensitive part of the fluorescence spectroscopy is the time-resolved spectroscopy [3]. In the time-resolved fluorescence spectroscopy, molecules are pumped with energy using a very short pulse of light with the length from pico- to nanosecond. It makes possible to watch the time evolution of the excited molecular population by observing the emission of photons in real time. The emission of a photon is a statistical process. Therefore, the time when an excided molecule remains in the excited state is also a statistical quantity. However, in an ensemble of identical molecules observed the decay statistics is well defined. In the simplest case, the number of molecules in the excited state decreases exponentially after exciting an ensemble of identical molecules by a short pulse:

$$n(t) = n(0) \cdot \exp(-t/\tau)\,, \tag{1}$$

where n is the number of molecules in the excited state, τ – lifetime of fluorescence. The eq. 1 can be written in the terms of fluorescence intensity. In this case it leads to a mono-exponential decay for the system with a single-type non-interacting fluorescent molecules:

$$F(t) = -\frac{dn(t)}{dt} = \frac{n(0)}{\tau}\exp(-t/\tau) = F_0 \exp(-t/\tau)\,. \tag{2}$$

If the system contains N types of the non-interacting fluorescent molecules, the eq. 2 transforms into a multi-exponential decay, presented in fig. 2*a* (F_0 was normalised to 1)

$$F(t) = \sum_{I=1}^{N} a_i \exp(-t/\tau_i), \quad \sum_{i=1}^{N} a_i = 1\,. \tag{3}$$

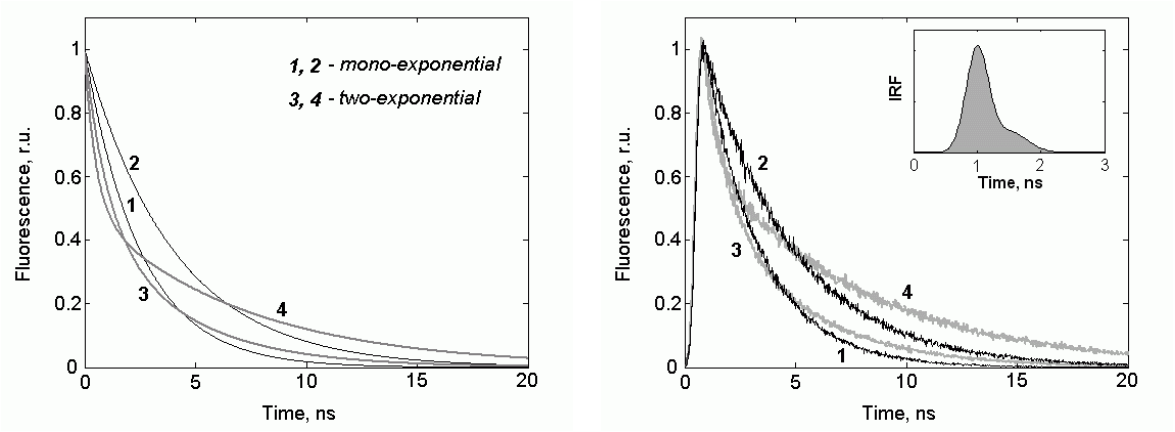
The interaction between fluorescent molecules in the form of quenching or non-radiative energy transfer leads to a significant complication of the decay law. In this case the fluorescence can be presented as a stretched exponent:

$$F(t) = F_0 e^{-ct} e^{-G(t)}\,, \tag{4}$$

where c is a constant and $G(t)$ is a geometry function which describes the distribution of fluorescent molecules. There is no analytical expression for $G(t)$ that would describe the behaviour of this system in the general case – it was derived only for several homogeneous systems and specific situations (uniform distributions of molecules in 1, 2, 3 dimensions).

The instrumental inaccuracy leads to additional complications in data analysis. The observed fluorescence is in fact the convolution between theoretical fluorescence (eq. 2 – 4) and the impulse response function (IRF) $h(t)$, which includes time lags of the light source pulse, photo-detector, optics and electronics. This gives us the real observed fluorescence (see fig. 2*b*) in the form

$$f(t) = h(t) \otimes F(t)\,. \tag{5}$$



a) *b*)

Fig. 2. . Theoretical (a) and experimental (b) fluorescence decays. The parameters of decays are the following: for the 1st curve τ=2.5 ns; for the 2nd curve τ=4 ns; for the 3d curve τ_1=4 ns, τ_2=1 ns;and for the 4th curve τ_1=7 ns, τ_2=0.5 ns.

Selection of a model for fluorescence data. The step of model selection is of a crucial importance for the analysis of fluorescence decays. The wrong model leads to complete loss of useful information.

In this section, we aim to the analysis of number of exponents in the multi-exponential decays. For example let ANN distinguish one-, two- and three-exponential decays. The three-layer perceptron was applied to recognise the model. Convoluted and distorted data, similar to those given in fig. 2*b*, were quantised within 32 bins of constant size. This information was given to 32 inputs of the three-layer perceptron with 16 neurons in two hidden layers. Number of outputs was equal to the number of recognised model (three models in this example). The value of the output shows some sort of "probability" to have the corresponding model for the analysed data.

The numerical experiment was performed to determine the accuracy of the approach. The single exponential model was determined correctly with the probability of 98%, for two-exponential model – 94% and for three-exponential model – 92%. Therefore, ANNs are able to perform the task of model recognition. The same technique can be used to distinguish between exponential and stretched-exponential decays.

Fluorescence spectroscopy: initial estimation of parameters. Consider the task of parameter estimation on the same three-exponential model. A three-layer perceptron with 16 neurons in each hidden layer was used to extract information about the lifetimes of exponential components. The scheme is similar to what we had for model selection. Pre-processed decays are given to the input of ANN and estimated and scaled lifetimes are taken from the outputs, therefore the number of inputs corresponds to the number of estimated parameters. Because of the form of eq. 3 the order of summation does not play any role. This may lead to the uncertainty during the training of ANN – same inputs may give different outputs. To avoid it the output values were artificially ordered: it was predetermined that longest lifetime is taken from the first output. The average and the shortest lifetimes are taken from the second and from the third output correspondingly.

The results of numerical experiments prove that ANN is able to perform the initial estimation of lifetimes in the case of multi-exponential decay. The probabilities of the estimation with 10% and 20% error were calculated. These probabilities are given in the table 1. In our numerical experiments we used simulated fluorescent decay.

Table 1. The probabilities to get estimation within the admissible error region

Admissible error interval	Probability to get τ_1 inside the admissible error region	Probability to get τ_2 inside the admissible error region	Probability to get τ_3 inside the admissible error region
< 10%	99%	73%	70%
< 20%	100%	97%	96%

It can be seen, that the ANN was able to estimate parameters of the three-exponential decays with a high probability. It should be mentioned that this method can be applied for more complex models (eq. 4) and it works directly with convoluted data achieving a high stability to noises.

4. Data analysis in ESR spectroscopy

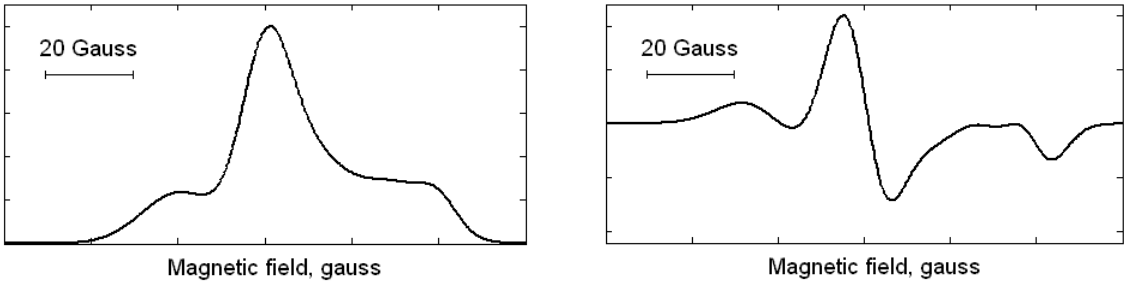
ESR spectroscopy. ESR spectroscopy exploits the physical phenomenon of absorption of microwave radiation by paramagnetic molecules or ions exposed to an external magnetic field. It is based on transitions between energy levels produced by an external magnetic field on an unpaired electron. This transition is detected as ESR signal or spectrum. The splitting of electronic energy fields in a magnetic field is used to determine structures of samples containing unpaired electrons.

In combination with labelling based on nitroxide spin probes, ESR spectroscopy is especially suitable for studying cell membranes. It can detect alterations caused by biologically active substances and indicating pathological conditions, such as acute phase, cancer, etc [2]. In general, it is used for the determination of mobility of molecules and their parts, and can provide some information about the molecular geometry in the region of 5-25 Å.

In the past, interpretation of ESR spectra was performed manually by measuring spectrum peaks characteristics and analyzing their relationships. However, the recorded ESR spectra provide much more reliable and biologically meaningful information when characterized through computer-aided spectrum simulation [4].

Experimental data. The shape of an ESR spectrum is strongly dependent on the mobility and orientation of a spin probe. The simplest types of spectra correspond to two opposite situations: probes are fixed in a crystal, having the same orientation, and fast non-restricted motions of probes. In these cases spectra can be presented by a three peak structure with Gaussian peaks for a fixed system and Lorentzian peaks for free one. Spectra that are much more complex are detected the situation when the motion of spin probe is slow or partially restricted. Each of three dimensional directions (x,y,z) gives its own spectrum, and the resulted one is called a powder spectrum (see fig. 3*a*). Most ESR spectrometers are designed to detect a slope of it (fig. 3*b*), and experimental data are always presented in the form of the spectrum derivatives. Derivative spectra gives spectra characteristics in a more visible form.

Specific simulation-optimisation techniques should be implemented to analyse ESR spectra [4]. A lot of information can be extracted from the positions of peaks in a powder spectrum. Below the ANN analysis *via* approximation is proposed. This method works with spectra themselves, not with their derivatives.



a) *b*)

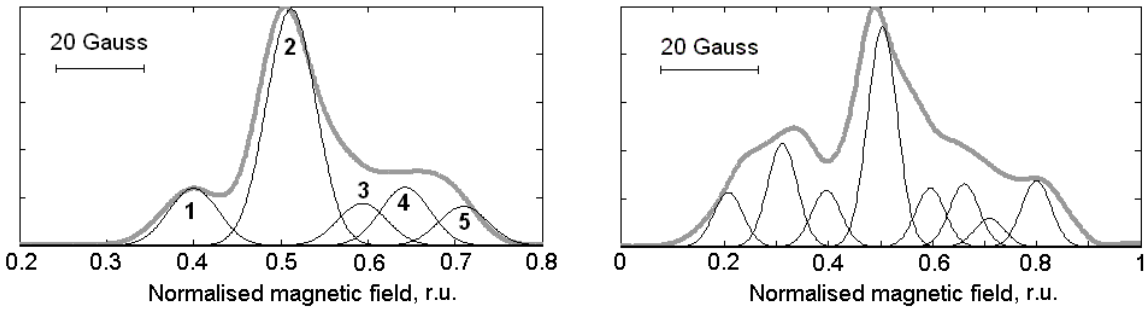
Fig. 3. ESR line shape for a powder spectrum (a) and its derivative (b)

Neural network analysis. To analyse an ESR powder spectrum and to study its interfering peaks, it is possible to approximate a spectral region by a set of analytical functions featuring physical lines. For data analysis we offer to use a special class of ANN, called radial basis function (RBF) networks [1]. RBF networks are known due to their ability to build a good approximation of experimental data by the mean of weighted sum of Gaussian or Lorentzian functions.

To keep approximation physically valid, a special criterion was included into the algorithm of ANN training. After each iteration, i.e. adding of a new neuron, the network was checked for the appearance of RBF with negative amplitude, and if one was found – the training stopped.

The number of numerical experiments was conducted to define the possibilities of RBF approximation. The results for one of those are given in fig. 4*a* and table 2. RBF-network accurately determined the number of peaks in simulated data. Their locations were determined with the error varying from 0 to 7%. To decrease the error and fit peaks precisely standard methods of multi parameter optimization can be used.

The experimental spectrum (fig. 4*b*) was recorded for *Mycoplasma* cells and lipids with incorporated 12NS probe, bound to bovine serum albumin aqueous medium [5].



a) *b*)

Fig. 4. ANN approximation of the simulated ESR powder spectrum (a) and the experimental one (b). Thick gray line shows the initial ESR spectrum, thin black curves present the estimated locations and heights of peaks

Table 2. Estimated locations of peaks for the simulated spectrum

Number of a peak	**1**	**2**	**3**	**4**	**5**
Real position (relative un.)	0.40	0.50	0.55	0.65	0.70
Estimated position	0.40	0.51	0.59	0.64	0.71
Relative error	0 %	2 %	7 %	2 %	1 %

5. Conclusions

The proposed approach was tested on the simulated fluorescence decays and ESR spectra. It showed rather good results in prediction of the model for fluorescence data. For the case of multi-exponential fluorescence decay analysis, the mean probability to obtain the correct lifetime values within the error range of ±10% was approximately 80%. It should be noted that the method is applicable in the case of non-exponential decays. The method can work with convoluted data. The absence of the deconvolution procedure gives a significant increase to its stability to noises.

The ANN with radial basis functions was successfully applied to approximate ESR powder spectra and to extract the positions of peaks. The method of RBF approximation is not restricted to Gaussian shape and may be applied to ESR spectra analysis in the case of a fast molecular motion (Lorentzian shape). The precision of the method can be increased by using of optimisation methods for the fine-tuning of peak parameters.

Authors would like to express their thanks to Dr. Marcus A. Hemminga (Laboratory of Biophysics, Wageningen University, The Netherlands) and Dr. Mikalai M. Yatskou (Belarusian State University, Minsk) for their kind assistance and information about experimental methods.

This work has been performed with a financial support of the Belarusian Republic Fund of Fundamental Researches (project "Development of methods and algorithms of statistical analysis of biomolecular systems").

References

[1]. M. Bishop *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press, 1997.

[2]. M. A. Hemminga, "Interpretation of ERS and saturation transfer ESR spectra of spin labeled lipids and membranes", *Chem. Phys. Lipids*, **32**, (1983), p. 323-383.

[3]. J. R. Lakowicz, *Principles of fluorescence spectroscopy*, Kluwer Academic/Plenum Publishers: New York, 1999.

[4]. B. Filipic, J. Strancar "Tuning EPR spectral parameters with a genetic algorithm", *Appl. Sot Comp.*, **1**, (2001), p. 83-90.

[5]. M. E. Tourtellotte, D. Branton, and A. Keith. "Membrane structure: spin labeling and freeze-etching of Mycoplasma laidlawii" *Proceedings of the National Academy of Sciences USA*, **66**, (1970), p. 909-916.