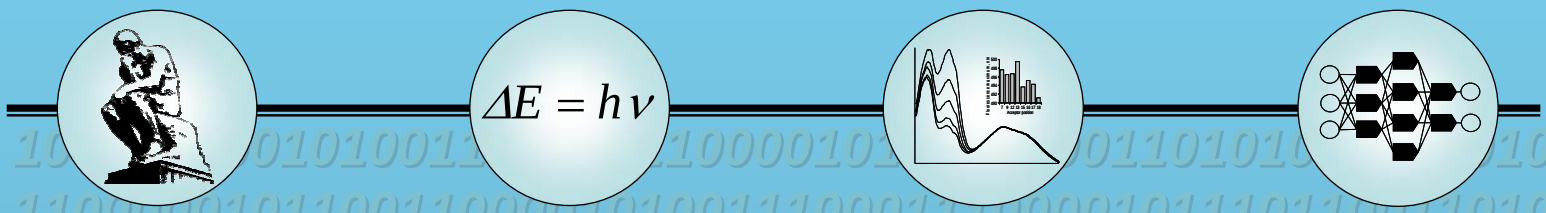
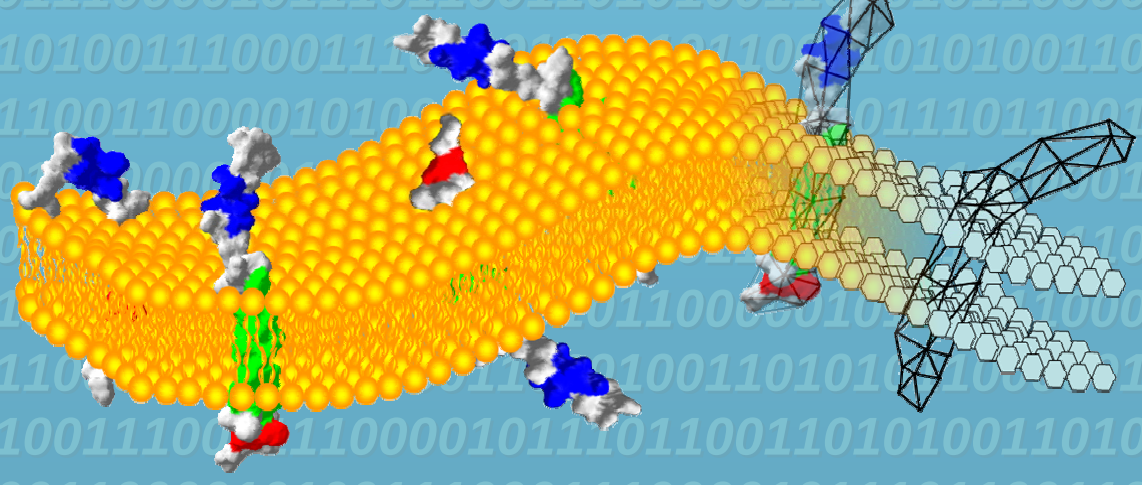
Simulation and Analysis of FRET in the Study of Membrane Proteins

Petr V. Nazarov





SIMULATION AND ANALYSIS OF FRET IN THE STUDY OF MEMBRANE PROTEINS

Petr V. Nazarov

Promotor:

Prof. Dr. H. van Amerongen,
Hoogleraar in de Biofysica
Wageningen Universiteit

Co-promotoren:

Dr. M. A. Hemminga,
Universitair Hoofddocent, Laboratorium voor Biofysica
Wageningen Universiteit

Prof. Dr. Vladimir V. Apanasovich,
Head of the Department of Systems Analysis
Belarusian State University

Promotiecommissie:

Prof. Dr. A. J. W. G. Visser	(Wageningen Universiteit)
Prof. Dr. J. A. Killian	(Universiteit Utrecht)
Prof. Dr. G. J. Fleer	(Wageningen Universiteit)
Prof. Dr. J. M. Vlak	(Wageningen Universiteit)

Dit onderzoek uitgevoerd binnen onderzoekschool EPS

SIMULATION AND ANALYSIS OF FRET IN THE STUDY OF MEMBRANE PROTEINS

Petr V. Nazarov

Proefschrift

ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit
Prof. dr. M. J. Kropff
in het openbaar te verdedigen
op woensdag 13 december 2006
des namiddags te half twee uur in de Aula

Nazarov, P. V.
Simulation and Analysis of FRET in the Study of Membrane Proteins,
Thesis, Wageningen University, 2006.

ISBN: 90-8504-553-3

To my mother, wife and son
for their love, support and understanding

CONTENTS

ABBREVIATIONS

AEDANS – N-(iodoacetylaminoethyl)-5-naphthylamine-1-sulfonic acid;

ANN – artificial neural networks;

DOPC – 1,2-Dioleoyl-sn-glycero-3-phosphocholine;

DOPG – 1,2-Dioleoyl-sn-glycero-3-[phosphor-rac-(1-glycerol)];

E. coli – Escherichia coli;

ESR – electron spin resonance;

FRET – Förster (or fluorescence) resonance energy transfer;

MLP – multilayer perceptron;

NMR – nuclear magnetic resonance

RBF – radial basis function;

SBF – simulation-based fitting.

1. GENERAL INTRODUCTION

1.1. Simulation as a tool for complex systems study

1.1.1. Types of models

Modeling is an essential part of all kinds of scientific studies of real-world objects. Being complex and many-sided, real objects cannot be investigated in all their manifestations; therefore we are forced to limit ourselves to study only a part of their properties. It means that from the very beginning of the study, a researcher builds in his mind a model of the object, which contains only its essential and interesting properties.

A model, associated with the studied object or system, can be a physical prototype (as in the case of physical modeling) or a formal system of concepts and relations (mathematical modeling), describing the object and its behavior with the required level of details.

Mathematical models can be divided into groups, based on the following criteria (Low and Kelton, 2000):

- Method of formal description (one can distinguish analytical and simulation models);
- Usage of time concept (static and dynamic models);
- Presence of stochastic components and relations (deterministic and stochastic models);
- Continuity (continuous and discrete models).

If the relations, which form a mathematical model, are rather simple and can be described using mathematical analytical expressions, analytical modeling can be used. The advantages of analytical modeling are its universality (in relation to the tasks of its application) and high precision. Unfortunately, the use of analytical models is not always possible. Systems of high complexity are currently studied in many science disciplines (informatics, electronics, astrophysics, biology, chemistry, economy). These systems contain a significant number of interacting components (which can be systems as well), diversity of interconnections, have a non-linear behavior, and, as a result, they are difficult to describe and predict. Usually an analytical prediction of such a system is concerned with a number of approximations and simplifications, resulting in rather rough estimations.

In such a case simulation modeling can be used instead of analytical modeling. Its idea can be reduced to two stages:

1. Building of a formal mathematical model, up to the complexity level at which an analytical description can be used.

2. Developing of a simulation algorithm, which imitates the system behavior, taking into account external influences and interactions between components, and followed by the realization of the algorithms as a computer program.

Using simulation modeling it is possible to study systems of almost any complexity. To develop a simulation model of a system, it is enough to know the partial behavior of its elements and to be able to simulate elementary interactions between them. Moreover, analytical models often operate with integral characteristics, which have no physical meaning. On the contrary, in simulation modeling the vast majority of the parameters has a physical meaning.

1.1.2. Simulation tasks during the study of complex systems

The tasks, to be solved by a researcher when studying complex systems can be roughly divided into two groups: direct and inverse tasks.

Let us understand under a direct task the prediction of the behavior of a system based on a known internal state and under different external conditions. This task can be easily solved by simulation modeling. To predict the system behavior, one has to run the simulation taking into account the initial conditions of the system and external influences. Stochastic and chaotic systems can be characterized in this way by statistical parameters obtained from simulations (therefore a number of independent simulation runs are essential). Examples of direct tasks are: molecular dynamics simulation, weather forecast, nuclear tests *in silico*, etc.

To solve the inverse task, means to find the hidden internal parameters of the system, or to define the structural model of a system by using external observations of the system behavior (experimental data). Examples of an inverse task are: analysis of experimental data, optimal design of a system, etc. The solution of the inverse task is much more complex than the direct task. For real systems the Kolmogorov's correctness condition (unique existence) (Kolmogorov, 1946) does not always hold and the observed data contain noise or are of a stochastic nature. At the same time in practice it is not necessary to find an ideal and exact solution. It is often enough to find a good and most probable estimation of the parameters sought.

1.1.3. Parameters identification using simulation modeling

Simulation-based fitting (SBF) approach was developed for the determination of physical parameters of complex systems, which cannot by described by analytical equations. The general scheme of the inverse task solution by simulation modeling is given in Fig. 1.1.

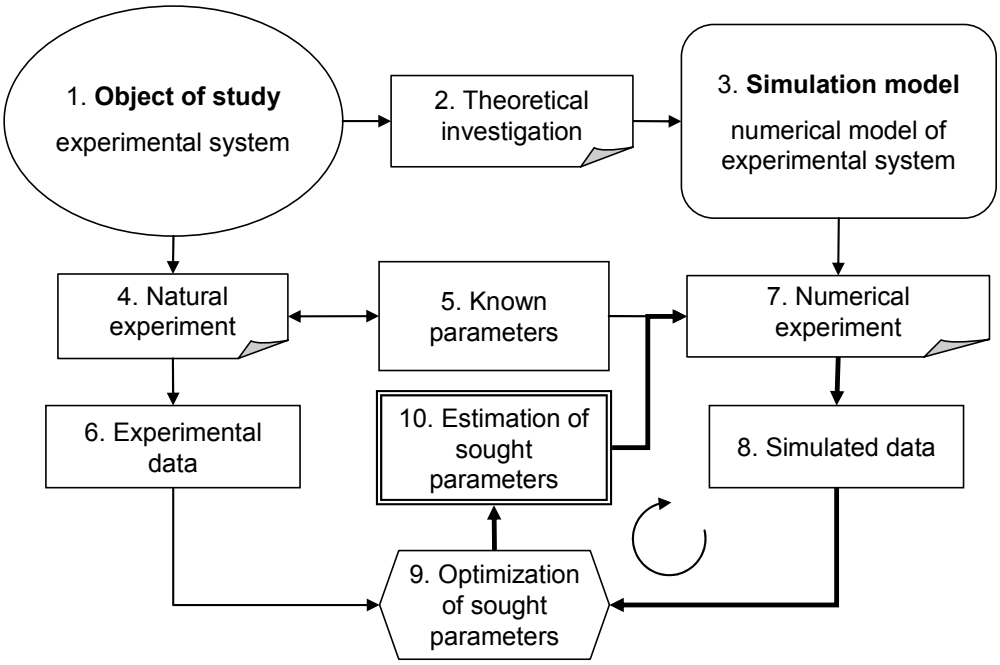


Figure 1.1. General solution of an inverse problem by the simulation-based fitting approach. Thick lines show the loop of the iterative estimation of the parameters sought.

For an experimental object (block 1) using theoretical investigation approaches (block 2) the simulation model (block 3) is built. Then a number of experiments is performed on the object (block 4) which results in a set of experimental data, block 6 (for instance – fluorescence spectra). The researcher knows the experimental conditions and this gives additional knowledge about the properties of the object (temperature, pH, viscosity, etc) – block 5. Using the known parameters of the object and making assumptions about the sought parameters (block 10), the researcher runs the simulation modeling and obtains simulated data (block 8) in this numerical procedure (block 7). Comparing simulated data and experimental ones (by calculation of the error function), the optimization procedure (block 9) changes the estimation of the sought parameters, and the algorithm cyclically goes to the numerical calculation again. This iterative search provides an estimation of the experimental parameters. If the model is adequate, and the inverse task is correctly formulated, the estimation found can be considered as a valid one.

The approach described above has some drawbacks:

- Significant time cost (computational expenses) – simulation modeling is usually concerned with a rather high computational load;
- The existence of a number of local minima of the error function, which necessitates multiple runs of the SBF with different initial estimations of the parameters sought.

To solve these problems special methods are used, for example neural network simulation ("black box" model) (Nazarov et al., 2004), parallel calculations (Fox et al., 1994; Nazarov et al., 2002), and multipoint optimization algorithms (genetic or evolutionary algorithms) (Tomassini, 1999). Some of these approaches will be considered in relation to the current work.

1.2. Artificial neural networks

1.2.1. Foundation of neural networks theory

An artificial neural network (ANN) consists of a number of simple processor elements, called artificial neurons. A classical artificial neuron includes a set of input weights w_i, summation unit, and an activation function. It can have a number of inputs, but only one output. The value of each input of a neuron x_i is multiplied by its adjustable weight coefficient w_i. One of the inputs, which always is present in a neuron, has a constant value $x_0 \equiv 1$. Its weight coefficient (w_0) is called the threshold of a neuron. A bounded function of infinite domain is applied to the weighted and summed inputs to limit the amplitude of the output signal y, as shown in Fig. 1.2 (Wasserman, 1989).

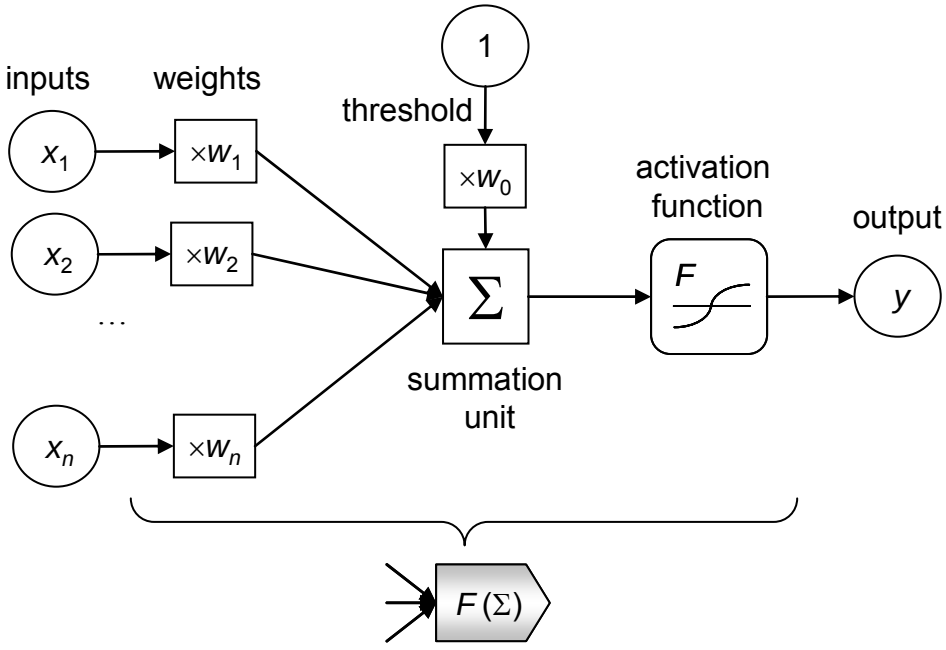


Figure 1.2. Classical artificial neuron with n inputs (x_i) and weight factors (w_i), threshold (w_0), non-linear activation function F and the symbolic notation of an artificial neuron (at the bottom).

Thus, the output of each neuron is calculated as

$$y = F\left(\sum_{i=0}^{n} x_i w_i\right) \tag{1.1}$$

We denote the sum of the weighted inputs $x_i w_i$ as Σ.

There are a number of different activation functions: threshold function, which is equal to 0 for $\Sigma \leq 0$ and 1 for $\Sigma > 0$, simple linear function, sigmoid function, hyperbolic tangent, and radial basis functions (for example Gaussian). The threshold function is similar to the nonlinear activation function of biological neurons (Wasserman, 1989). However, this function does not have a continuous derivative, and therefore is not often used in complex ANNs. The sigmoid function represented by Eq. 1.2 allows compressing possible values of Σ into the interval (0, 1). This function can be differentiated, and this property is widely used in training algorithms.

$$y = \frac{1}{1 + e^{-\Sigma}} \tag{1.2}$$

A classical ANN is build of artificial neurons as is depicted in Fig. 1.3. The information is given to the inputs X_i of the ANN and then propagates inside the ANN between the embedded neurons. If an ANN has feedback connections (see the dotted lines in Fig. 1.3) this propagation can be infinite in time (this property is used in such applications of ANN as digital filtering, control, dynamical simulation, etc.) (Kolen and Kremer, 2001).

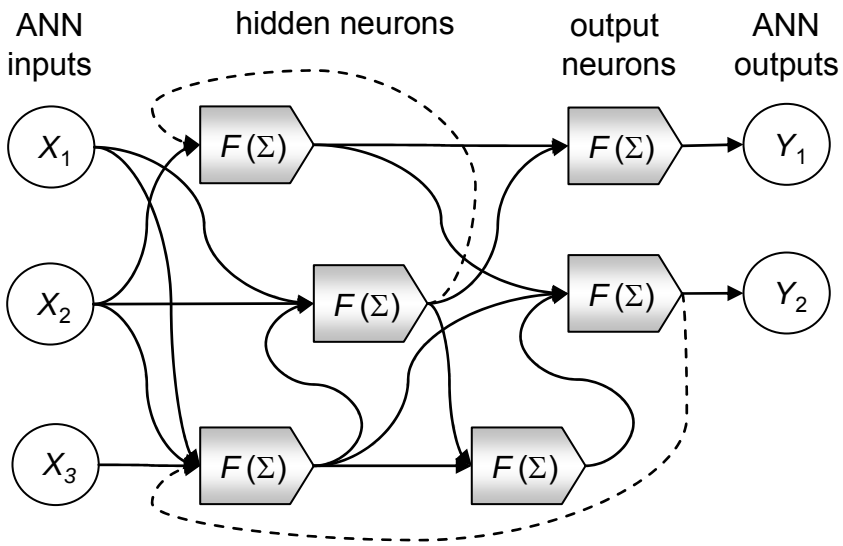


Figure 1.3. ANN with three inputs (X_i) and two outputs (Y_j). Feedback connections are marked by dotted lines.

The functionality of an ANN is provided by a proper selection of the adjustable parameters w_i for each neuron in the network. The selection and adjustment process is called "training" of the ANN. There are two general approaches for the training of an ANN – supervised and unsupervised (adaptive) training strategies.

During supervised training a special data set (training set) is used, containing input values and network target (desired) output values. Each element of this set, which consists of the inputs and outputs of the ANN, is called a training pair (Stegemann and Buenfeld, 1999). Of course, a single training pair is not very informative, therefore a representative set of training pairs should be used to train a neural network. During training the inputs from the training pairs are provided to the inputs of the ANN and the outputs are calculated. Based on the deviation between the obtained and target outputs, the weights of the ANN are modified in compliance with the training algorithm.

In the second, unsupervised training approach (or self-organization), the ANN organizes the training data and discovers its emergent collective properties without training pairs. In unsupervised training, the network is provided with inputs but not with desired outputs. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaptation. At present, unsupervised learning is not well understood. This adaptation to the environment in principle can allow an ANN to continually learn on its own. The best known representatives of ANN's which utilize this approach are Kohonen's networks or self-organized maps (Kohonen, 1984).

1.2.2. Multilayer perceptron

In this work a special class of ANN called multilayer perceptron (MLP) is used. Let us therefore consider this type of neural networks in detail. A multilayer perceptron is an ANN in which the neurons are located in layers, and the outputs of the neurons of the k-th layer are connected only with the inputs of the neurons of layer k-1. This structure is shown in Fig. 1.4. The absence of feedback and lateral (inside one layer) connections results in a fast unidirectional propagation of information from inputs to outputs of the MLP.

MLP has a rather good and well studied method of supervised training, called *back propagation error*, which is a kind of gradient minimization algorithm adapted for the case of the high dimension of the parametric space of the ANN (Hagan and Menhaj, 1994). Each training pair contains a vector $\boldsymbol{X}_T$ of ANN inputs and a vector $\boldsymbol{Y}_T$ of ANN target outputs. The general scheme of the training algorithm is the following:

1. The network is initialized by giving random values (usually around ± 0.2) to all adjustable weight coefficients and thresholds.
2. For each training pair $(\boldsymbol{X}_T,\boldsymbol{Y}_T)_k$ put $\boldsymbol{X}_T$ into the network input and calculate its output vector $\boldsymbol{Y}$.
3. Calculate the error – the discrepancy between calculated $\boldsymbol{Y}$ and desired $\boldsymbol{Y}_T$.
4. Correct the weights of the output layer based on the gradient method to decrease the error.
5. Correct the weights of the hidden layers (layer by layer).
6. Repeat steps 2–5 until the stopping criterion is met.

In the current work the MLP, being a universal approximator, was used for "black box" modeling and replacement of a simulation model (see Chapter 4). The variable parameters of the simulation model are put to the inputs of the MLP and the simulation results are taken from its outputs. The selection of the MLP configuration is discussed in Chapter 4.

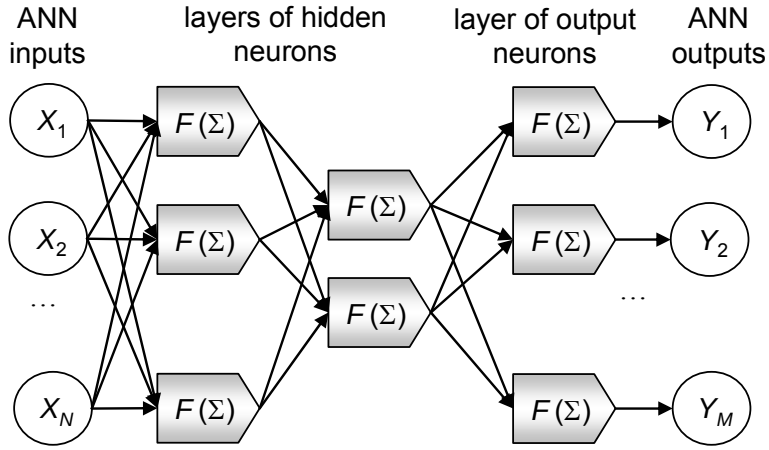


Figure 1.4. Example of a multilayer perceptron with N inputs (X_i), M outputs (Y_j) and three layers.

1.3. Membranes and membrane proteins

1.3.1. Biological membranes and their experimental models

Lipid membranes play a key role in the structural organization and functionality of all cells. Membranes form intracellular compartments, and thus separate the content of a compartment from its environment. However this is not the only function of a membrane, they also regulate all interactions between compartments. Examples of such regulations are physical ion and molecule transport, and information transfer, due to conformation changes of

membrane components. Moreover, many cell enzymes are located at membrane surfaces. Some of them catalyze transmembrane reactions even when reagents are separated by the membrane. Other enzymes form complexes, which perform a chain of consecutive chemical reactions and structure transformations, and their efficiency is increasing due to the fact that the enzymes are located in a two-dimensional space. Membranes also play a role in important biological processes, such as, replication of prokaryotic DNA, protein biosynthesis, protein secretion, bioenergetic processes, etc. (Gennis, 1989).

Biological membranes can be rather complex (Fig. 1.5 A) and include a lot of biomolecules of different nature, for example:

- different types of lipids (Fig. 1.5 B): phospholipids, sterols, etc;
- membrane proteins, glycoproteins, etc.

The main building elements of membranes are lipids (mainly phospholipids) (Gennis, 1989) – organic molecules, which have hydrophilic "heads" and hydrophobic "tails" (Fig. 1.5 B). This feature of the lipids results in the formation of various structures when they are mixed with water. Some of these structures are presented in Fig. 1.5 C–E.

The formation of such structures is a very interesting example of self-organization of biomolecules. The formation of a specific structure is determined by a number of physicochemical factors, such as: type and concentration of lipids, temperature, pH, salts, mechanical way of mixing, etc.

The most interesting model systems, imitating biological membranes, are large unilamellar vesicles (Fig. 1.5 D). Having rather big, approximately 50-500 nm diameter and hence small curvature, these liposomes can be considered as almost flat lipid bilayers in which membrane proteins and other membrane components are present in their original conformations. Moreover, the water suspension of large unilamellar vesicles has a high optical transparency, which is an important condition for fluorescent experiments as used in the present work.

1.3.2. General characteristics of membrane proteins

Proteins (polypeptides) are the most important type of biomolecules, which particularly allow the existence of life and participate in all cellular processes as biological regulators (enzymes), transport systems (pores), transport paths and cytoskeleton (actin), nanomachines aimed at assembling complex macromolecules, etc (Apell and Karlish, 2001; Byrne and Iwata, 2002; Pollard and Borisy, 2003; Stryer, 1978; Torres et al., 2003).

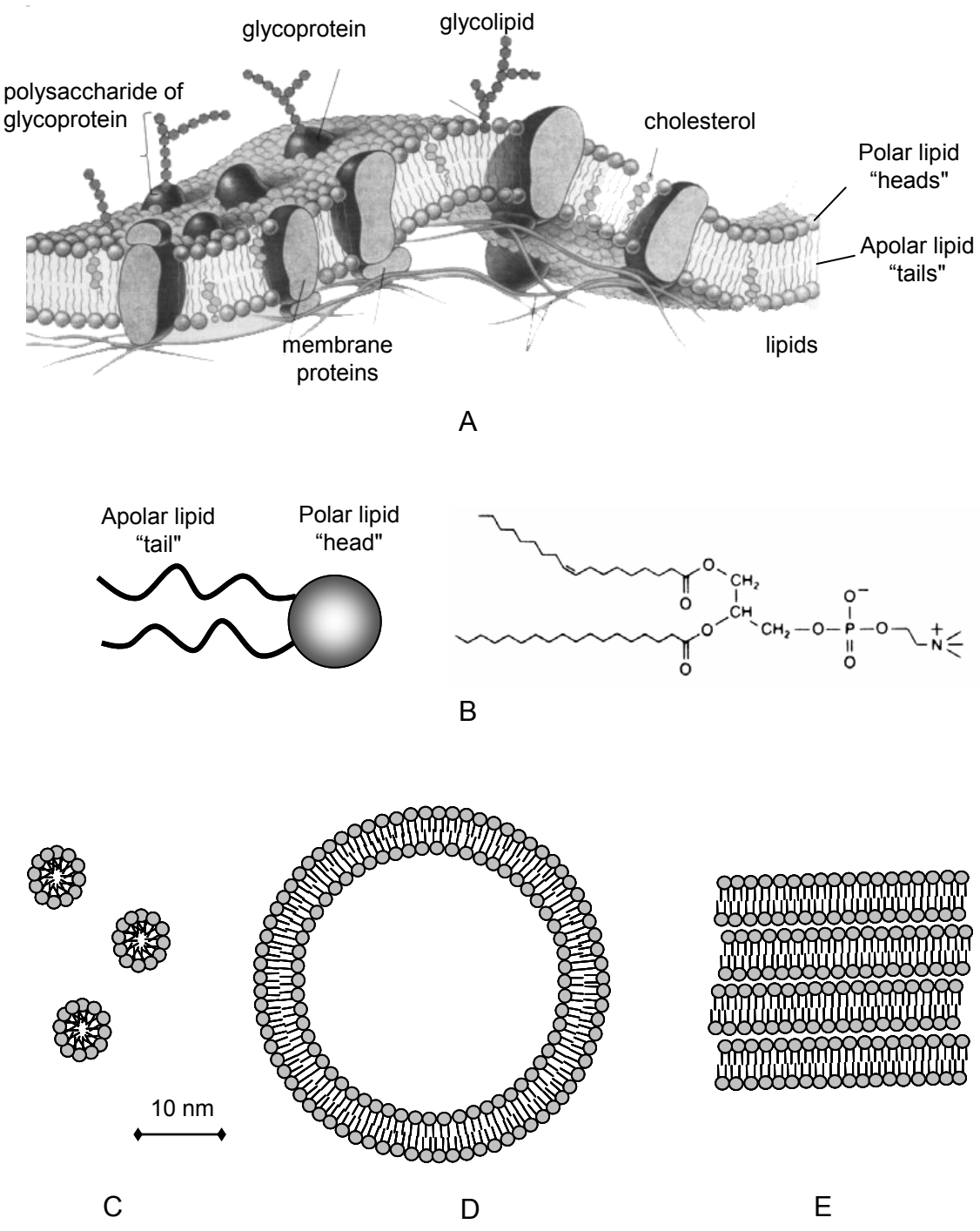


Figure 1.5. Biological membrane (A) (Chiras, 2002), chemical structure of phospholipids (B) and several structures formed by lipids when mixed with water: micelles (C), vesicle (unilamellar liposome) (D) and oriented bilayers (E).

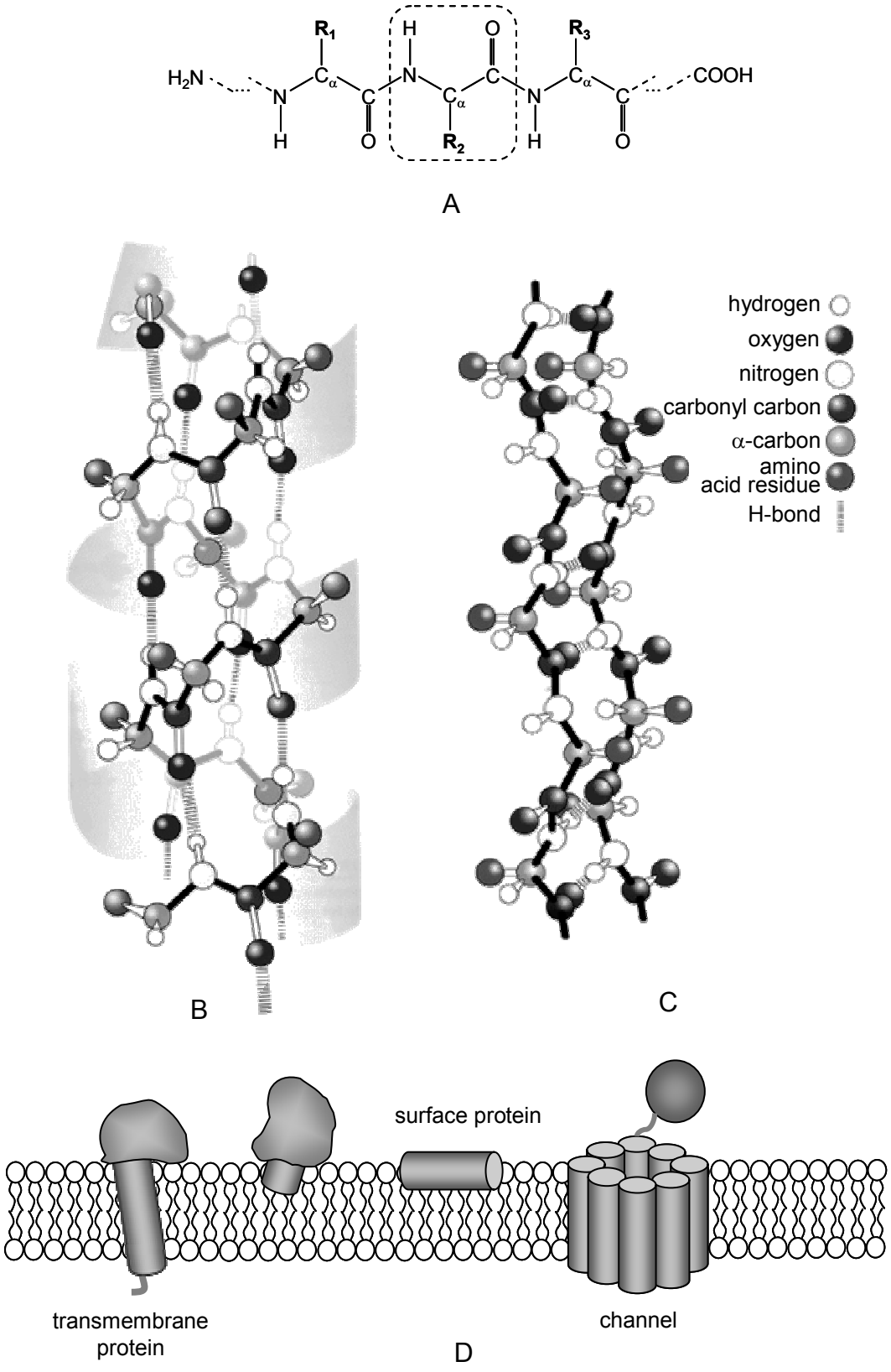


Figure 1.6. Polypeptide link in a protein (A), and two ordered secondary structures of proteins: α-helix (B) and β-sheet (C) conformations (Devlin, 2006). Different types of membrane protein incorporation into a membrane (D).

Being biopolymers, proteins are built from 20 different amino acids. All amino acids have the structure and the type of interconnection depicted in Fig. 1.6 A. Among the main properties of amino acids, which determine the structure and functions of a protein, the following can be listed: physical size and shape of an amino acid residue, polarity, charge, presence of sulfur able to make sulfur bridges (via cysteines).

The amino acid sequence of a protein is called the primary protein structure. Depending on this structure and polarity of the environment hydrogen bonds between the hydrogen of NH group and the oxygen of the C=O group may occur, which determine the secondary structure. One example of a secondary structure is the α-helix, which is considered in this work for building protein structural models (Fig. 1.6 B). This helix has a vertical step of 0.15 nm per 1 amino acid, and a complete turn of the helix contains 3.6 amino acids. Another wide spread secondary structure is the β-sheet (Fig. 1.6 C), in which two closely situated protein chains form mutual hydrogen bonds.

Biological membranes contain 20 – 80% of proteins (by weight). Usually membrane proteins are responsible for functional activity of membranes. Different enzymes, transport proteins, receptors, channels, pores, etc., which are all representatives of membrane proteins, provide the unique functions of each membrane. The position of membrane proteins in lipid bilayers is quite diverse and results from the polarity of the amino acid residues (Fig. 1.6 D). For example, membrane proteins can pass through a membrane (transmembrane proteins), or lie on top of a bilayer (surface membrane proteins) (Gennis, 1989).

The study of membrane proteins, even at the level of primary structure, was significantly impeded at the initial stage of research, because of their bad solubility in water. Today this problem has been successfully solved, due to newly developed solvents (Gennis, 1989); however, their study still is a challenging area in structural biochemistry.

1.3.3. Complexity of membrane protein studies

Despite the importance of the study of membrane proteins both for general understanding of cellar processes and for new drugs development, up to now no universal methods have been developed, able to determine their structure in a general case. About 20 000 protein structures are known today and less than 1% of them correspond to membrane proteins (Arora and Tamm, 2001; Torres et al., 2003; White, 2004). The up-to-date data

related to the structures of membrane proteins obtained are published on the web page of Stephen White's group[1].

The standard approach to protein structure determination is X-ray crystallography. This method is extremely precise and provides information at the atomic-scale level (resolution up to 0.05 nm). The method analyses X-ray diffraction patterns obtained with protein crystals. The obvious limiting condition is the ability to prepare a protein crystal. Unfortunately, the preparation of crystals of membrane proteins is an extremely complex process. Furthermore, one should be aware of the fact that the structure of a membrane protein in its crystal form may not be the same as in the membrane-bound form (dos Remedios and Moens, 1995; Hemminga et al., 1993; Torres et al., 2003).

The second widespread method of membrane protein structure determination is nuclear magnetic resonance (NMR) spectroscopy. This method is applicable, as a rule, to small and average-size membrane proteins and allows finding distances between atoms, if they are in the range of 0.1-0.5 nm. Two different experimental approaches are used for the study of membrane proteins. In liquid NMR spectroscopy membrane proteins are studied in micelles (Fig. 1.5 C) (Papavoine et al., 1998; Papavoine et al., 1997). The protein-containing micelles should have a relatively low molecular weight (< about 25 000 Dalton). At the same time the small sizes of micelles may lead to distortions of the original protein structures and introduce excessive dynamics (Vos et al., 2005). Another approach is the application of solid-state NMR to study the protein-lipid system in parallel oriented dehydrated bilayers (Fig. 1.5 E). This approach gives perfect results in the case when the studied protein is completely buried in the lipid bilayer and has an insignificant polar part. However, if the protein is situated simultaneously in the bilayer and in the outer water environment, the lack of water layer between lipid bilayers (~ 0.4 nm) may again lead to distortion of its structure (Vos et al., 2005).

The drawbacks mentioned above stimulated us to search for alternative approaches to study membrane proteins. Such methods are electron spin resonance spectroscopy (ESR) (Meijer et al., 2001b; Stopar et al., 2002) and fluorescence spectroscopy. An important example of the latter is Förster resonance energy transfer (FRET) spectroscopy, which is based on dipole-dipole interactions between fluorophores (Förster, 1965) and able to determine intra- and intermolecular distances (Lakowicz, 1999). This method allows obtaining distances in the range of 1-10 nm and therefore it can be used not only for

[1] http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html

membrane protein structure determination, but also for the study of protein-protein interactions (Fernandes et al., 2003; Lakey et al., 1993; Li et al., 1999; Stryer, 1978).

The present work aims at the development of FRET methods and related experimental data analysis, to extract relevant structural information about membrane proteins. As a test system for this study, the well-known bacteriophage M13 major coat protein is used.

1.3.4. Major coat protein of bacteriophage M13

Bacteriophage M13 is a small filamentous virus of *Escherichia coli*. A single viral particle has a diameter about 6.5 nm and contains 2800 copies of a coat protein, which prevents its single-stranded DNA from damage (Marvin and Hohn, 1969; Stopar et al., 2002). The length of viral particles varies with the length of the DNA. The main part (98%) of the viral coat, which represents a hollow flexible cylinder with thickness 1.5-2 nm, consists of one type of protein, called M13 major coat protein.

During the penetration of bacteriophage into *E. coli*, its coat is diluted in the bacterial cell membrane and the DNA gets into the cell. Infected *E. coli* cells start the replication of viral DNA, which stimulates the production of new viral proteins. Coat proteins are incorporated into the outer membrane where they stay until the exit of new viral particles (Stopar et al., 2002). Assembling of new viral particles occurs in the bacterial membrane with viral and host proteins participating in this process. Interestingly, the exit of new viral particles is not followed by lysis or any other damage of the outer membrane. Infected *E. coli* continues the production of viral particles during its lifetime. The scheme of viral invasion and new viral particles exit is described in detail in (Hemminga et al., 1993). This allows to get an almost unlimited quantity of viral proteins from an infected culture. Moreover, it allows to produce large quantities of mutants of coat proteins by mutations in viral DNA (Sambrook et al., 1989).

After integration in the *E. coli* membrane, the viral major coat protein adopts a transmembrane configuration. The protein consists of 50 amino acid residues (Fig. 1.7 A) and has a hydrophobic region (positions 21-39). In this figure the colors represent the polarity of the amino acid residues, in accordance to the scale of White and Wimley (White and Wimley, 1999).

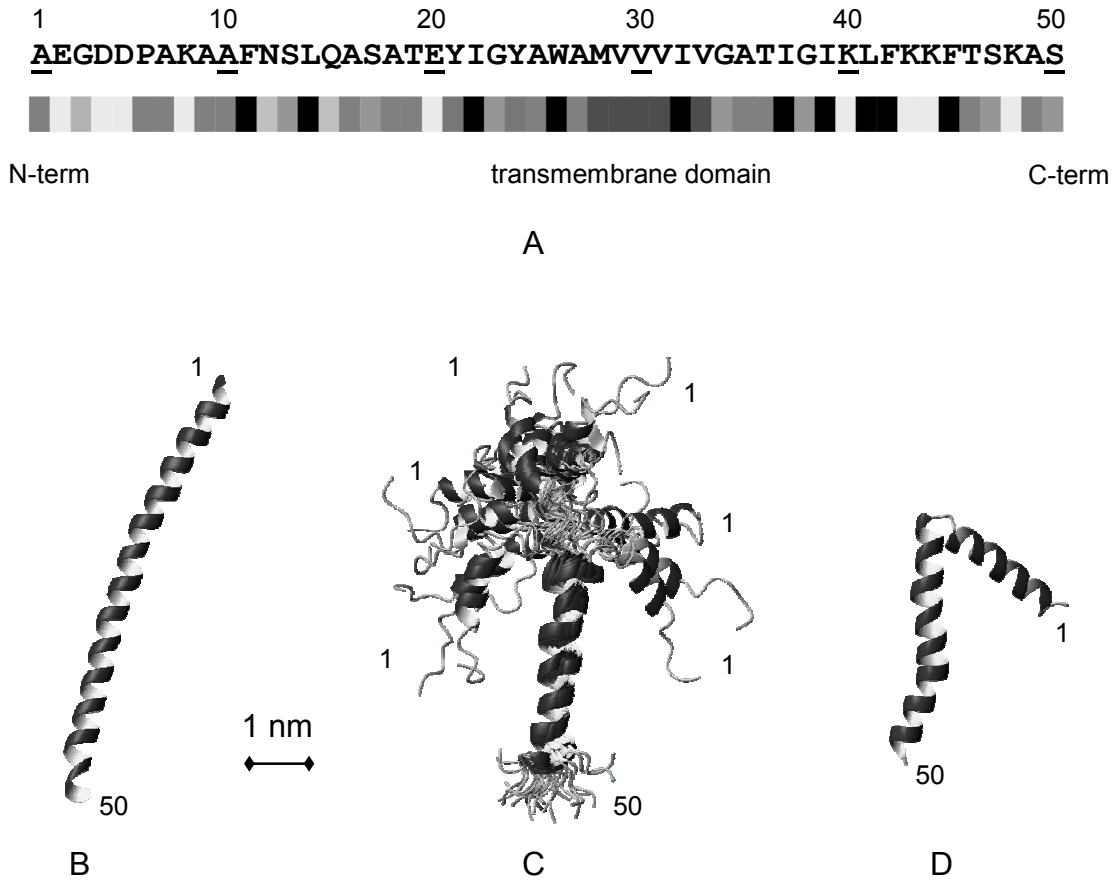


Figure 1.7. Major coat protein of bacteriophage M13. Amino acid sequence of the protein and polarity of amino acid residues (based on the scale of White and Wimley (White and Wimley, 1999)): a darker color corresponds to a more hydrophobic residue (A). Protein structures determined in: viral particle (Marvin, 1998) (B), lipid micelles (Papavoine et al., 1998) (C), and dehydrated bilayers (Marassi and Opella, 2003) (D).

The knowledge of its primary structure and the possibility of site-directed mutagenesis has made this protein an excellent test and reference system to study fundamental protein-lipid interactions and its structure when embedded in a membrane. Unfortunately, despite a lot of studies on this subject, the discussions about the precise structure of M13 coat protein still continue. It is known that in the viral particle the protein represents a slightly curved α-helix (Fig. 1.7 B), as was found by X-ray crystallography (Marvin, 1998). This structure is very stable, as it is necessary to protect the viral DNA from external influences. In detergent micelles (Fig. 1.5 C) by using NMR (Papavoine et al., 1998), however, the protein is found to be in a very dynamic state (Fig. 1.7 C). However, as was mentioned before, micelles have small sizes, and their high curvature can distort the original structure of membrane proteins (Vos et al., 2005). Other authors, who studied the coat protein in a dehydrated bilayer (Marassi and Opella, 2003) found a rigid L-shape structure (Fig. 1.7 D). As was mentioned in (Vos et al., 2005), dehydrated bilayers may also introduce artifacts, because of a lack of water

layer between the lipid bilayers (approximately 0.4 nm). Therefore, it seems worthwhile to study the M13 major coat protein in vesicles, which most closely mimic a biological membrane.

Several studies have been devoted to the study of the dynamics and membrane embedment of M13 protein using ESR spectroscopy (Meijer et al., 2001b; Meijer et al., 2001a; Stopar et al., 2002; Stopar et al., 2005; Stopar et al., 2006b). As a result of these studies it was found that this protein most probably consists of two stable α-helices at amino acid positions 7-16 and 25-45. The second α-helix represents the transmembrane protein domain and the first one is partially situated in water, partially – in the lipid head group region.

The protein has also been studied with fluorescence spectroscopy. The work of Fernandes et al. (Fernandes et al., 2004; Fernandes et al., 2003) was aimed on protein lipid interactions and protein-protein associations. As a result it was qualitatively determined that the protein does not show a tendency to aggregate and its distribution in DOPC:DOPG bilayers can be considered as uniformly random. Koehorst et al. were the first to study the position and orientation of M13 coat protein in bilayer using fluorescence spectroscopy (Koehorst et al., 2004).

1.4. Photophysical processes and fluorescence spectroscopy

1.4.1. The method of fluorescence spectroscopy

Fluorescence spectroscopy is a powerful tool, which allows the study of the structure and dynamics of molecular systems. Examples of experimental systems, successfully studied with fluorescence techniques are: polymers, solutions of surfactant species, thin films, biomembranes, proteins, nucleic acids, and living cells. The diversity of physical parameters, which can be characterized by fluorescence, is astonishing: intra- and intermolecular distances, polarity, viscosity, structural ordering, molecular mobility, and electric potential (Valeur, 2001). One of the main advantages of the fluorescence method is its high sensitivity. For instance, modern techniques allow the tracking of a single fluorescent molecule (Keller et al., 1996).

Fluorescent probes (fluorophores) are needed to perform fluorescence experiments. These fluorophores can either be introduced artificially into the system under investigation, or can be originally present (for example, the fluorescent amino acid tryptophan is widely used in protein studies).

In this section I will consider some basic principles of photophysical processes occurring in experimental systems when studied by fluorescence methods. Special attention will be given to the Förster resonance energy transfer (FRET) technique. In addition, the phenomenon of energy transfer by dipole-dipole resonance interaction will be considered in detail.

1.4.2. Resonance energy transfer

Consider a simplified diagram of energy levels (Jablonski's diagram) for two fluorophores (called donor and acceptor, see Fig. 1.8 A). When the donor absorbs a photon of a proper wavelength it goes to one of the excited states (S_1 or S_2 in Fig. 1.8 A). The absorption of a photon is a very fast process, which takes ~10^{-15} s. Then a fast relaxation to the lowest exited state level S_1 occurs. This process is called internal conversion (in the figure its rate is noted as k_{IC}) and the corresponding time scale is ~10^{-12} s. Because the usual lifetime of the S_1 state is in the range from 10^{-9} to 10^{-8} s, internal conversion occurs before other processes take place, such as relaxation to S_0, or energy transfer (Lakowicz, 1999).

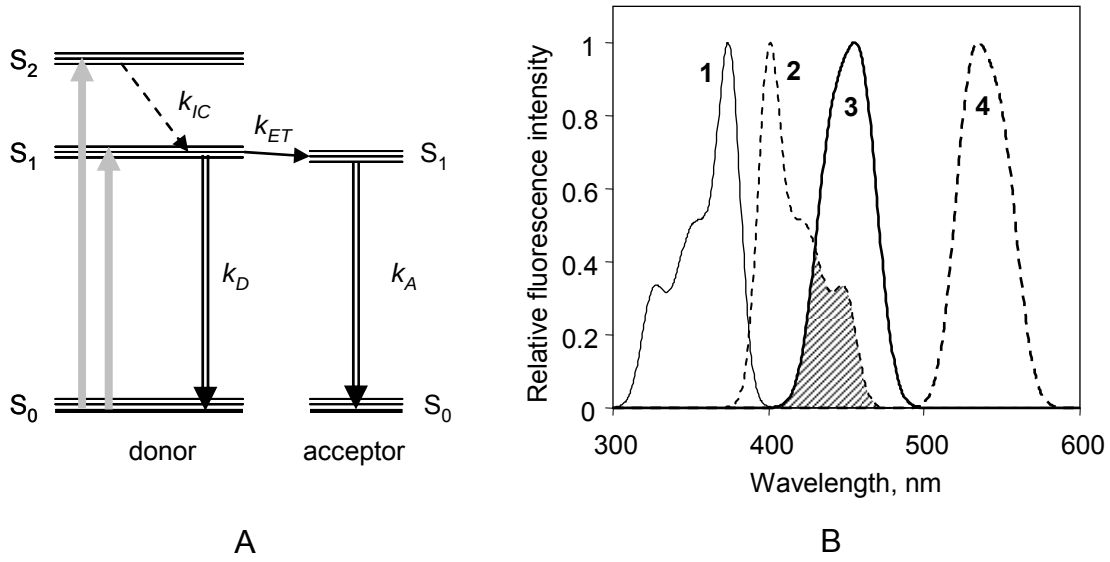


Figure 1.8. Simplified scheme of energy levels (Jablonski's diagram) for the donor-acceptor pair (A) and schematic normalized spectra for donor and acceptor (B). In figure (A) S_0 denotes the ground states of donor and acceptor, $S_{1,2}$ – excited states, k_{IC} corresponds to the rate of internal conversion, k_D, k_A – donor and acceptor sum relaxation rates, k_{ET} – the rate of energy transfer. In figure (B) curves 1 and 2 represent the absorption and emission spectra of the donor; 3 and 4 are the absorption and emission spectra of the acceptor. The donor-acceptor spectral overlap area is indicated via hatching.

An isolated fluorescent molecule, being in the S_1 state can relax to S_0 by emission of a photon (fluorescence) or via radiationless relaxation. In the figure the sum rates of relaxation for isolated fluorophores are denoted as k_D and k_A. In the case of isolated non-interaction molecules these rates are in inverse proportion to the life times in the excited states τ_D and τ_A (Lakowicz, 1999).

When two fluorescent molecules (donor in excited state and acceptor in ground state) are close to each other and their spectra overlap as is shown in Fig. 1.8 B, the probability of donor-acceptor energy transfer exists. Consider an interaction between these two molecules from the quantum physics point of view. Let us assume for simplicity, that only one electron from the donor and one electron from the acceptor are participating in such an interaction. The wave function for the initial state (only the donor is excited) Ψ_0 and the final sate (only the acceptor is excited) Ψ_1 can be written in the following form (Valeur, 2001):

$$\begin{aligned} \Psi_0 &= \frac{1}{\sqrt{2}}\left(\Psi_{D^*}(1)\Psi_A(2) - \Psi_{D^*}(2)\Psi_A(1)\right) \\ \Psi_1 &= \frac{1}{\sqrt{2}}\left(\Psi_D(1)\Psi_{A^*}(2) - \Psi_D(2)\Psi_{A^*}(1)\right) \end{aligned}, \tag{1.3}$$

where wave functions of the electron of the donor and acceptor are marked by *D* and A, respectively. The numbers 1 and 2 mark the electron (electron 1 initially is located near the donor, and 2 near the acceptor), and the asterisk denotes the excited state.

The interaction matrix element, describing the coupling between the initial and final states can be written as

$$U = \left\langle \Psi_0 \middle| V \middle| \Psi_1 \right\rangle, \tag{1.4}$$

where V is the perturbation part of the total Hamiltonian of the system $\hat{H} = \hat{H}_D + \hat{H}_A + V$. The interaction given by Eq. 1.4 can be written as a sum of two components

$$\begin{aligned} U &= \left\langle \Psi_{D^*}(1)\Psi_A(2) \middle| V \middle| \Psi_D(1)\Psi_{A^*}(2) \right\rangle - \left\langle \Psi_{D^*}(1)\Psi_A(2) \middle| V \middle| \Psi_D(2)\Psi_{A^*}(1) \right\rangle = \\ &= U_c - U_{ex} \end{aligned} \tag{1.5}$$

The first component U_c, characterizes the Columbic interaction between the multipoles. This component results in a transition of the electron of the donor to the ground state with a simultaneous transition of electron of the acceptor to the excited state. The second component U_{ex}, describes the physical exchange of electrons between the donor and acceptor.

This interaction is a quantum-mechanical phenomenon caused by the symmetry of the wave functions with respect to the spin and coordinates exchange for the two electrons.

The Coulombic interaction can be divided into dipole-dipole, dipole-quadrupole, quadrupole-quadrupole, and other types of multipole-multipole interactions. However in the vast majority of cases a first-order approximation is used, taking into account only dipole-dipole interaction U_{dd} between the emission dipole moment of the donor $\boldsymbol{M}_D$ and the excitation dipole moment of the acceptor $\boldsymbol{M}_A$ (Förster, 1948; Lakowicz, 1999; Valeur, 2001). The dipole-dipole compound of the interaction can be written in the following form

$$U_{dd} = \frac{\boldsymbol{M}_D \cdot \boldsymbol{M}_A}{r^3} - 3\frac{(\boldsymbol{M}_A \cdot \boldsymbol{r})(\boldsymbol{M}_D \cdot \boldsymbol{r})}{r^5}, \tag{1.6}$$

where $\boldsymbol{r}$ is the vector connecting the centers of donor and acceptor. Expression (1.6) can be rewritten if the angles between the vectors are taken into account:

$$U_{dd} = a\frac{|\boldsymbol{M}_D||\boldsymbol{M}_A|}{r^3}(\cos\theta_{DA} - 3\cos\theta_D\cos\theta_A), \tag{1.7}$$

Here a is a coefficient that depends on the selection of the measuring system (for example, if U_{dd} is in cm^{-1}, the distance is in nm, and the moments are in debye, $a = 5.04$), θ_{DA} is the angle between the dipoles, and θ_D, θ_A are the angles between vector $\boldsymbol{r}$ and the dipoles of the donor and acceptor, respectively. The value of U_{dd} can be significant up to a distance of 8-10 nm (dos Remedios and Moens, 1995; Lakowicz, 1999; Valeur, 2001).

The theory that describes this mechanism of energy transfer was developed in detail by Förster and published in 1948 (Förster, 1948). It should be mentioned, that the dipole-dipole approximation is valid for the case of point dipoles, i.e. when the donor-acceptor distance is much larger than the physical size of the molecule groups. The distance range where this theory is applicable is approximately 1-10 nm.

The second contribution in Eq. 1.5, describes the exchange mechanism (U_{ex}). This means a physical electron exchange and is possible only when the electron clouds are overlapping, in other words when physical contact between the donor and acceptor occurs. Therefore this contribution is relevant only at short distances, because the electron density decays exponentially outside a molecule. For two electrons at a distance r_{12} the spatial part of the interaction can be written as:

$$U_{ex} = \left\langle \Phi_{D*}(1)\Phi_A(2)\left|\frac{e^2}{r_{12}}\right|\Phi_D(2)\Phi_{A*}(1)\right\rangle, \quad (1.8)$$

where Φ_D, Φ_A – are the contributions of the spatial wave functions into the total wave functions Ψ_D, Ψ_A (that includes spin functions as well).

The energy transfer rate, mentioned in the beginning of this section, can be obtained for each type of interactions using Fermi's Golden Rule:

$$k_{ET} = \frac{2\pi}{\hbar}U^2\rho, \quad (1.9)$$

where ρ is a measure of the density of the interacting initial and final states, as determined by the Franck-Condon factor and is related to the overlap integral between the emission spectrum of the donor and the absorption spectrum of the acceptor (Valeur, 2001).

By substituting Eqs. 1.7 and 1.8 into Eq. 1.9, it is possible to obtain expressions for the rate constants of dipole-dipole and exchange mechanisms.

In the case of dipole-dipole interaction for the isolated donor-acceptor pair the rate constant of energy transfer can be written as

$$k_{ET}^{dd} = k_D\left(\frac{R_0}{r}\right)^6 = \frac{1}{\tau_D}\left(\frac{R_0}{r}\right)^6. \quad (1.10)$$

Here we assume that the donor-acceptor distance r is constant during the donor lifetime in the excited state. R_0 is the so-called Förster distance, which is equal to the distance at which excitation can be transferred to the acceptor with probability 0.5, i.e. when $k_{ET} = k_D$. This constant is calculated using spectroscopic data for the participating fluorophores. The expression for R_0 is as follows

$$R_0^6 = \frac{9000(\ln 10)\kappa^2 Q_D}{128\pi^5 N_A n^4}\int_0^\infty F_D(\lambda)\varepsilon_A(\lambda)\lambda^4 d\lambda \quad (1.11)$$

In this expression Q_D is the quantum yield of the isolated donor (i.e. the probability of fluorescence after donor excitation), N_A is the Avogadro constant, n is the refractive index of the donor-acceptor intervening medium (Knox and van Amerongen, 2002), F_D the normalized area of the donor emission spectrum, ε_A is the acceptor molar extinction coefficient, and κ^2 is an orientation factor, dependent on the directions of the transition dipoles as in (1.7):

$$\kappa = \cos\theta_{DA} - 3\cos\theta_D \cos\theta_A . \quad (1.12)$$

The orientation factor κ^2 is the most not obvious parameter of Eq. 1.9 (Dale et al., 1979). It can vary from 0 to 4 (collinear dipoles). A dynamical averaging of κ^2 has to be taken into account, if the speed of dipole reorientation is significant in comparison with the donor life time. For sufficiently fast isotropic rotation of dipoles this results in $\kappa^2 = 2/3$ (Dale et al., 1979). In many experimental works it is shown that even a small mobility of fluorophores (when it is accompanied by random orientation of moments) allows the use of the value of $\kappa^2 = 2/3$ (dos Remedios and Moens, 1995; Kamal and Behere, 2002; Lakshmikanth et al., 2001; Loura et al., 1996).

One can obtain the exchange rate by substituting Eq. 1.8 into 1.9:

$$k_{ET}^{ex} = \frac{2\pi}{h} K \exp\left(-\frac{2r}{L}\right) \int_0^\infty F_D(\lambda)\varepsilon_A^*(\lambda)d\lambda , \quad (1.13)$$

where r is the donor-acceptor distance, L the average Bohr radius (0.1–0.2 nm for aromatic molecules), ε_A^* the normalized absorption spectrum of the acceptor, and K is a constant not related to spectroscopic properties. Because it is hard to determine K experimentally, it is hard to use the exchange effect for quantitative data analysis (Valeur, 2001).

1.4.3. Resonance energy transfer in case of an ensemble of molecules

In our work only one of above mentioned mechanisms of energy transfer will be taken into account, namely dipole-dipole. This mechanism often prevails under experimental conditions in the study of biomolecular systems, such as proteins, DNA, and membranes (dos Remedios and Moens, 1995; Lakowicz, 1999; Loura et al., 1996). Mainly photosynthetic complexes, for which much more complex excitonic energy transfer models are valid, can be considered as exceptions to this rule (van Amerongen et al., 2000). Therefore, in the following, k_{ET} will reflect the rate constant for the dipole-dipole compound of resonance energy transfer.

One of the important characteristics of energy transfer is energy transfer efficiency. Physically, energy transfer efficiency is the mean probability of energy transfer from a donor to an acceptor. For an isolated donor-acceptor pair, separated by distance r, the efficiency can easily be calculated using relaxation rates:

$$E = \frac{k_{ET}}{k_D + k_{ET}} = \frac{(R_0/r)^6}{1+(R_0/r)^6} = \frac{1}{1+(r/R_0)^6}. \quad (1.14)$$

If the fluorescence of an isolated donor f_D represents a mono-exponential decay with rate $k_D = \tau_D^{-1}$, as shown in Eq. 1.15, the fluorescence of the donor in a donor-acceptor pair is mono-exponential as well, and has the same initial intensity F_D, see Eq. 1.16.

$$f_D(t) = F_D \exp[-k_D t] \quad (1.15)$$

$$f_{DA}(t) = F_D \exp(-k_D t - k_{ET} t) = F_D \exp\left[-k_D\left(1+(R_0/r)^6\right)t\right] \quad (1.16)$$

If the system contains an ensemble of donors and acceptors, for each donor all pathways of deexcitation should be taken into account. Therefore the energy transfer efficiency can be calculated as an average for all donors using the expression

$$E = \langle E_i \rangle_{N_D} = \left\langle \frac{\sum_{j=1}^{N_A} (R_0/r_{ij})^6}{1+\sum_{j=1}^{N_A} (R_0/r_{ij})^6} \right\rangle_{N_D}. \quad (1.17)$$

Here index i reflects donor enumeration, j the acceptor enumeration, N_D, N_A are the numbers of donors and acceptors in the system, respectively, and $\langle ... \rangle_{N_D}$ indicates averaging over the donor ensemble. Correspondingly, the precise expression for donor fluorescence is written as follows.

$$f_{DA}(t) = \frac{F_D}{N_D} \sum_{i=1}^{N_D} \exp\left[-k_D\left(1+\sum_{j=1}^{N_A}(R_0/r_{ij})^6\right)t\right] \quad (1.18)$$

Obviously the application of Eqs. 1.17 and 1.18 is possible only for a relatively small number of fluorophores in the system. Therefore analytical models have been developed, which describe donor fluorescence in the presence of uniformly distributed acceptors (Blumen et al., 1986; Davenport et al., 1985; Dewey and Hammes, 1980). All analytical expressions are concerned with simplifications and approximations. In the model of Blumen et al. (Blumen et al., 1986), represented by Eq. 1.19, donor and acceptor molecules have infinite small sizes and are uniform randomly distributed in a d-dimensional space:

$$f_{DA}(t) = F_D \exp\left[-k_D t - \Gamma(1-d/6)\cdot\sigma_{dA}\cdot V_d \cdot R_0^d \cdot (k_D t)^{d/6}\right], \quad (1.19)$$

where d is the dimensionality of the system, Γ is the gamma function, σ_{dA} is the acceptor density for the dimensionality d (for planar systems – surface density, for the space – concentration), and V_d is the volume of a d-dimensional sphere of unitary radius (for a planar system $V_d = \pi$, for the space $V_d = 4\pi/3$).

Independently of this universal expression Davenport et al. (Davenport et al., 1985) proposed an analytical solution for a layered planar system (see Eq. 1.20). In this system donors and acceptors are distributed in two parallel planes separated by the distance h (identical molecules in the same plane). Donors have sizes, or a donor exclusion distance R_E, which means that R_E is the minimal distance between acceptors and the donor projection on the acceptor plane.

$$f_{DA}(t) = F_D \exp\left\{ -k_D t - 2\pi\sigma_A h^2 \int_{\alpha=0}^{\frac{h}{\sqrt{h^2+R_E^2}}} \left(1 - \exp\left[-k_D t \left(\frac{R_0 \alpha}{h} \right)^6 \right] \right) \alpha^{-3} d\alpha \right\}. \quad (1.20)$$

For both analytical solutions the efficiency of energy transfer is calculated using the following expression

$$E = 1 - \frac{\int_0^\infty f_{DA}(t)dt}{\int_0^\infty f_D(t)dt}, \quad (1.21)$$

which simplifies in the case of mono-exponential donor decay to

$$E = 1 - \frac{k_D}{F_D} \int_0^\infty f_{DA}(t)dt. \quad (1.22)$$

1.4.4. Methods of energy transfer efficiency determination

The simplest way of energy transfer efficiency determination is given by Eq. 1.21. During experiments the overall fluorescence intensity of the donors in presence (experimental sample) and absence (reference sample) of acceptors should be measured. If it is hard to keep the donor concentration constant in both samples, the following normalization should be used:

$$E = 1 - \frac{C_D}{C_{DA}} \frac{\int_0^\infty f_{DA}(t)dt}{\int_0^\infty f_D(t)dt}, \tag{1.23}$$

where C_D is the donor concentration in the reference sample, and C_{DA} the experimental one. The detected fluorescence of donors in the presence of acceptor f_{DA} should originate solely from donors and be free from acceptor fluorescence. This method is applicable for both steady-state and time-resolved fluorescence spectroscopy.

Another method, providing information about energy transfer efficiency, is based on the detection of acceptor excitation (Lakey et al., 1993). Consider steady-state fluorescence (emission) of acceptors F_A^λ, where λ is the excitation wavelength. Let us denote Θ_A as the part of the energy that is absorbed by the acceptor due to direct excitation and Θ_D the part of energy that is absorbed by donors and then transferred to acceptors. Then the excitation spectrum of an acceptor can be written as:

$$F_A^\lambda = \left(\Theta_A^\lambda + \Theta_D^\lambda E\right)\gamma, \tag{1.24}$$

where γ is some experimental factor, including the optical path of the beam in the sample, excitation radiation intensity, etc. By calculating the fluorescence of the sample at two wavelengths, which correspond to maximal donor λ_D and acceptor λ_A absorptions, it is possible to obtain the system of Eq. 1.25, originating from Eq. 1.24. The energy transfer efficiency E can easily be evaluated from it. In Eq. 1.25 Θ is given in terms of molar extinction coefficients ε of the donor and acceptor at different wavelengths multiplied by the concentrations C_D and C_A of donors and acceptors, respectively.

$$\begin{cases} F_A^{\lambda_D} = \left(C_A \varepsilon_A^{\lambda_D} + C_D \varepsilon_D^{\lambda_D} E\right)\gamma \\ F_A^{\lambda_A} = \left(C_A \varepsilon_A^{\lambda_A} + C_D \varepsilon_D^{\lambda_A} E\right)\gamma \end{cases}, \tag{1.25}$$

Often for real systems the absorption of the donor at the acceptor excitation wavelength is very small, thus $\varepsilon_D^{\lambda_A} \approx 0$. In this case the expression for the energy transfer efficiency, calculated from Eq. 1.25 is

$$E = \frac{C_A}{C_D}\left(\frac{F_A^{\lambda_D}}{F_A^{\lambda_A}} - \frac{\varepsilon_A^{\lambda_D}}{\varepsilon_A^{\lambda_A}}\right)\frac{\varepsilon_A^{\lambda_A}}{\varepsilon_D^{\lambda_D}}. \tag{1.26}$$

1.5. Application of FRET to protein studies

From the very beginning, FRET was aimed first of all to distance determination in proteins and their complexes. The theory of FRET was initially developed and its applicability overviewed by Förster (Förster, 1948, Förster, 1965). Later, Stryer and Haughland (Stryer and Haugland, 1967) studied, amongst others, synthetic polypeptides labeled with naphthyl and dansyl. The authors experimentally demonstrated the theoretical dependency between energy transfer efficiency and donor-acceptor distance (see Eq. 1.14). In the next review of Stryer (Stryer, 1978) the wide possibilities of FRET spectroscopy for protein structure determination were demonstrated. Also the classical works of Vanderkooi (Vanderkooi, 2002; Vanderkooi et al., 1977) were aimed at fluorescent and phosphorescent methods for protein investigation.

Most of the FRET applications can be conditionally divided into three groups. The first one is related to direct intramolecular distance determination between protein sites or between proteins in their complexes (Almeida and Opella, 1997; Lakshmikanth et al., 2001; Torres et al., 2003). The second application is related to the study of specific and non-specific binding of different molecular formations. The studies aimed at the determination of the spatial distribution of biomolecules can also be related to this group (Davenport et al., 1985; Fernandes et al., 2003; Förster, 1948). Finally, the third application is the study of macro objects (organelles, cells) by fluorescence microscopy. Usually, special luminescent proteins (yellow, green and cyan fluorescence proteins) are used for this. Such kind of FRET applications can, for instance, be found in (Rao and Mayor, 2005; Vogel et al., 2006).

The important advantage of FRET in the study of proteins is the almost complete control over amino acid sequence. Especially, this holds for small proteins, which can be synthesized artificially (Hesselink et al., 2005; Killian, 2003; Sparr et al., 2005), or proteins obtained from bacterial cultures (Spruijt et al., 1989). This allows site-specific modification of the amino acid sequence, and therefore site-directed labeling. The method of site-directed labeling can give a lot of useful structural information, especially in the study of membrane proteins.

1.6. Short overview of data analysis methods for fluorescence spectroscopy

As was mentioned before, initially the main application of FRET spectroscopy was the determination of intramolecular distances for donor-acceptor pairs. However, the application of expressions for analysis of energy transfer in more complex systems was hampered,

because of the complexity of the mathematical models for energy transfer. The main mathematical models at this stage were Eqs. 1.14 and 1.16. For this reason, in many studies only qualitative information was used.

In the 80s several groups published analytical equations, which describe energy transfer in three dimensions (Blumen et al., 1986) or two dimensions (Davenport et al., 1985; Dewey and Hammes, 1980) (see Eqs. 1.19, 1.20) in the previous section). These expressions are still used up to now (Fernandes et al., 2004; Loura et al., 1996) for analysis of intermolecular energy transfer. However, based on the investigation reported in this thesis, together with the work of Berney (Berney and Danuser, 2003), Eqs. 1.19 and 1.20 have to be used carefully, taking into account the approximations which have been made for their derivation (dimensionless acceptor molecules, uniform random distribution of fluorophores, low excitation intensity, absence of any interactions in the system except dipole-dipole coupling between donor and acceptor).

Quite often researchers do not build the detailed model of the processes in the system, and use an integral analytical description instead. In this case, for instance, the law of donor fluorescence decay given by Eq. 1.18 can be approximated by the sum of several exponents. In practice, for a rather precise description of any decay, it is sufficient to use 4-10 exponential components (Hiriayama et al., 1990). However this approximation gives only a rough estimation of energy transfer and can be used mainly for a qualitative description of photophysical processes.

From the 90s a new powerful tool became available that enabled the study and analysis of photophysical processes: computer simulation. The application of computer simulation to FRET problems was developed in the papers by Andrews, Berberan-Santos and Demidov (Andrews and Demidov, 1999; Berberan-Santos and Valeur, 1991; Demidov and Borisov, 1993). Computer simulation does not only allow the analysis of complex molecular systems (Frederix et al., 2002; Yatskou et al., 2001a; Yatskou et al., 2003), but also provides the possibility to check the quality of previously developed analytical models (Berney and Danuser, 2003).

Speaking about methodological aspects of experimental data analysis, the method of global data analysis should be mentioned. The method applies a simultaneous analysis of all available experimental data by a single model. Partial quality factors are combined into a global criterion. This methodology was first applied to fluorescence data by Beechem et al. (Beechem and Brand, 1985; Beechem and Brand, 1986; Beechem and Haas, 1989). The effectiveness and stability of this approach was demonstrated in their papers. It allows

avoiding local minima (or at least decreasing their number) during fitting and increases the noise stability. Global data analysis was used in the current work as well. The numerical tests performed on synthetic data confirm the effectiveness of the global analysis approach.

1.7. Outline of the Thesis

This thesis is devoted to the development of advanced methods for analysis of fluorescence data, based on simulation modeling, global analysis approach, and artificial neural networks. Especially the advantages and problems of the simulation-based fitting (SBF) approach for fluorescence data analysis are considered. The methods and algorithms developed are applied in particular to study the structure and embedment of membrane proteins in artificial membranes.

In Chapter 2 the methodologies of global analysis and SBF are applied to obtain information about the position and orientation of M13 major coat protein in DOPC:DOPG vesicles. The spatial model for the protein-lipid system is described and a full mathematical description of the energy transfer processes in the studied system is presented. Furthermore an algorithm for the analysis of SBF solution sets is provided. The resulting physical parameters that describe the embedment and orientation of the protein in the membrane, such as protein-protein aggregation, protein depth, tilt angle, and tilt direction, are in good accordance with previously reported values.

The methodology described in Chapter 2 is further extended in Chapter 3. In addition to the global analysis and SBF information was used, obtained from the fluorescence Stokes shift. A novel fuzzy “rules” approach was used to filter the resulting solutions, based on the position of the fluorescence maximum for different label positions. This analysis results in an improved structure of the M13 major coat protein that turns out to be close to a single helix from amino acid residue 10 to 50.

Chapter 4 is aimed at the enhancement of simulation-based fitting by an artificial neural networks (ANN) simulation. The main idea of the improvement is the replacement of the “white-box” simulation model, by a “black-box” neural-network model during the fitting procedure. The method was tested on the simulation model for Förster (or fluorescence) resonance energy transfer (FRET) data in a protein-lipid system. It was found that this method is a valid approach and can be applied when the number of variable (fitted or changed during experiments) parameters of the simulation model is less than (or in some cases equal to) 6. The method results in a considerable speeding up of the simulation (about a factor of 10^4).

Finally, Chapter 5 demonstrates the possibility of a direct application of ANNs to spectral analysis. In this chapter ANN was applied to spectra obtained in intracavity laser absorption spectroscopy. It was shown that ANN is able to deal quite well with stochastic noise and frequency-domain irregularities of the laser pulse and offers improved quality of the analysis.

2. FRET STUDY OF MEMBRANE PROTEINS: SIMULATION-BASED FITTING FOR ANALYSIS OF MEMBRANE PROTEIN EMBEDMENT AND ASSOCIATION

Petr V. Nazarov, Rob B.M. Koehorst, Werner L. Vos, Vladimir V. Apanasovich, Marcus A. Hemminga

Published in *Biophysical Journal*, **2006**, 91, p. 454-466.

ABSTRACT

A new formalism for the simultaneous determination of the membrane embedment and aggregation of membrane proteins is developed. This method is based on steady-state Förster (or fluorescence) resonance energy transfer (FRET) experiments on site-directed fluorescence labeled proteins in combination with global data analysis utilizing simulation-based fitting. The simulation of FRET was validated by a comparison with a known analytical solution for energy transfer in idealized membrane systems. The applicability of the simulation-based fitting approach was verified on simulated FRET data and then applied to determine the structural properties of the well-known major coat protein from bacteriophage M13 reconstituted into unilamellar DOPC:DOPG (4:1 mol/mol) vesicles. For our purpose, the cysteine mutants Y24C, G38C, and T46C of this protein were produced and specifically labeled with the fluorescence label AEDANS. The energy transfer data from the natural tryptophan at position 26, which is used as a donor, to AEDANS were analyzed assuming a helix model for the transmembrane domain of the protein. As a result of the FRET data analysis the topology and bilayer embedment of this domain were quantitatively characterized. The resulting tilt of the transmembrane helix of the protein is 18 ± 2°. The tryptophan is located at a distance of 8.5 ± 0.5 Å from the membrane center. No specific aggregation of the protein was found. The methodology developed here is not limited to M13 major coat protein and can be used in principle to study the bilayer embedment of any small protein with a single transmembrane domain.

2.1. Introduction

Membrane proteins play an important role in almost all cell activities. They perform a staggering range of biological reactions including respiration, signal transfer, molecular and ion transport (Byrne and Iwata, 2002). However, the structure determination of membrane proteins is still at the frontier of structural biology. While 30-40% of all proteins are membrane proteins, yet less than 1% of the known protein structures are for membrane proteins (Arora and Tamm, 2001; Torres et al., 2003). (For the most recent state for membrane proteins of known structure, see: http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html.) The complexity and delicacy of membrane-protein systems substantially impede the application of standard methods of protein study, such as X-ray crystallography and NMR (dos Remedios and Moens, 1995; Torres et al., 2003). Furthermore, these techniques are aimed at short-range structural information, and seem to be not useful for the study of long-range interactions, for instance in the case of protein association and clustering.

These factors impel to find other approaches to study proteins incorporated into lipid bilayers. A successful alternative is Förster (or fluorescence) resonance energy transfer (FRET) spectroscopy (Förster, 1965; Lakowicz, 1999; Stryer, 1978). This technique provides distance information within a range of 10-100 Å, which is sufficient to study the structure of membrane proteins and their complexes. FRET spectroscopy has been successfully applied to several problems in biology as a means of estimating intra and intermolecular distances in macromolecular systems, especially proteins (Lakey et al., 1993; Li et al., 1999; Stryer, 1978). The idea of FRET is labeling of the macromolecules with fluorescent labels of two kinds – a donor and an acceptor, and analysis of radiationless dipole-dipole energy transfer (Förster, 1948) between them. One of the advantages of such an approach is that several natural amino acid residues of a protein, such as Trp and Tyr, can be utilized as fluorescent labels (dos Remedios and Moens, 1995; Fleming et al., 1979).

Despite the elegant analytical models for a uniform planar donor-acceptor distribution that were developed two decades ago (Davenport et al., 1985; Dewey and Hammes, 1980; Wolber and Hudson, 1979), the complexity of protein-lipid systems hampers and limits an analytical interpretation of FRET data (Berney and Danuser, 2003; Frederix et al., 2002). For example, in the present work several numerical tests were performed to study the applicability of analytical models for the analysis of membrane protein systems. It was found that analytical expressions give incorrect results when the size of acceptor-host molecules is

comparable to the Förster distance of the donor-acceptor pair. On the other hand, simulation modeling of photophysical processes in an experimental system during a fluorescence measurement was proven to be a powerful alternative to analytical modeling, not restricted to special conditions (Berney and Danuser, 2003; Frederix et al., 2002; Nazarov et al., 2004; Yatskou et al., 2001a). The standard approaches to simulate FRET effects in complex systems are various Monte Carlo simulation schemes (Berney and Danuser, 2003; Frederix et al., 2002; Yatskou et al., 2001a). However, Monte Carlo simulation modeling is a very time-consuming operation. Furthermore, a time-resolved approach is not needed for the analysis of steady-state FRET data.

The goal of the current work is to develop and test a methodology for the analysis of steady-state FRET data to build a low-resolution structural model of a protein-membrane system with a quantitative characterization of its parameters. To perform this goal a steady-state FRET model is built and utilized in a simulation-based fitting (SBF) approach to approximate the experimental data by their simulated analogues (Nazarov et al., 2004; Yatskou et al., 2001a). By comparison with standard analytical data fitting techniques, simulation modeling has the advantage that it operates with the physical parameters of the system itself and gives a direct insight in how they affect the experimental characteristics of the system.

The methodology developed is tested on a well-known coat protein from bacteriophage M13. During a part of its life cycle, the coat protein is stored as a membrane protein in the *E. coli* host. Therefore it is an excellent model system to study fundamental aspects of protein-lipid and protein-protein interactions (Stopar et al., 2003). This single membrane-spanning protein consists of 50 amino acid residues and has mainly an α-helical conformation. The protein has been extensively studied in model membrane systems by several biophysical techniques (Fernandes et al., 2004; Fernandes et al., 2003; Glaubitz et al., 2000; Koehorst et al., 2004; Marassi and Opella, 2003; Meijer et al., 2001b; Papavoine et al., 1998; Papavoine et al., 1997; Spruijt et al., 2000; Stopar et al., 2002; Stopar et al., 2003; Vos et al., 2005). For FRET studies, the natural single tryptophan residue of the protein at position 26 (Trp-26) was used as a donor label. To introduce an acceptor label to the protein, a number of mutants, containing unique cysteine residues at specific positions, was produced. The cysteine residues were labeled with the fluorescent label N-(acetylaminoethyl)-5-naphthylamine-1-sulfonic acid (AEDANS) (Spruijt et al., 2000). This label was used as an acceptor. To separate intra and intermolecular energy transfer contributions, we performed titration experiments in which we added wild-type protein to mutant proteins at different L/P

ratios. Both unlabeled mutant and wild-type protein can be considered spectroscopically identical as donor-containing molecules without acceptor label. The labeled mutants contain both a donor and acceptor. In a fluorescence excitation experiment one can optically select the labeled mutant proteins by monitoring the acceptor fluorescence. In a fluorescence excitation spectrum FRET can be deduced from the enhancement of acceptor fluorescence at the donor absorption wavelength. Upon addition of donor containing wild-type protein the intermolecular energy transfer component is increased exclusively.

In this paper we focus our analysis on the transmembrane domain of the protein, which was recently found to be in an almost perfect α-helix conformation (Koehorst et al., 2004; Vos et al., 2005). To take into account the membrane embedment of the proteins and possible protein aggregation, a model of a protein-lipid bilayer system is generated. This model is then used in an SBF approach to analyze the fluorescence data. To make the SBF procedure more effective a global analysis strategy is applied, in which all data are analyzed simultaneously. This approach provides information about the membrane embedment of the transmembrane protein domain in terms of protein depth, tilt angle, and protein association.

2.2. Experimental

2.2.1. Sample preparation

The lipid bilayer systems were prepared from dioleoylphosphatidylcholine (DOPC, 18:1PC) and dioleoylphosphatidylglycerol (DOPG) lipids in a 4:1 molar ratio, denoted as DOPC:DOPG. DOPC was purchased from Avanti Polar Lipids and DOPG was purchased from Sigma.

Site-specific cysteine mutants of M13 major coat protein were prepared, purified and labeled with AEDANS (Molecular Probes) as described previously (Spruijt et al., 2000). Wild-type protein and AEDANS-labeled M13 coat protein mutants were reconstituted into phospholipid bilayers as reported earlier (Spruijt et al., 1989).

For this study we used AEDANS-labeled cysteine mutants of M13 coat protein with the cysteine residue at positions 24 (Y24C), 38 (G38C), and 46 (T46C). Titration experiments were performed in which the wild-type protein concentration was increased, whereas the mutant concentration was kept constant. The sample conditions for these titrations are given in Table 2.1. For the purpose of correcting the fluorescence results (see $\varepsilon_A^{290}/\varepsilon_A^{340}$ in Eq. 2.1), we also used a mutant (Y21A/Y24A/W26A/G23C) having the AEDANS labeled cysteine at position 23, in combination with a threefold mutation of the tryptophan at position 26 and the

tyrosines at positions 21 and 24 into alanines. The labeling efficiency of the mutants having the AEDANS label at position 24, 38, and 46 was determined as reported previously (Spruijt et al., 1996) and amounted to 62, 55, and 69%, respectively. The labeling efficiency is explicitly taken into account in Table 2.1 in the ratio of the number of unlabeled to labeled proteins (r_{ul}), as it affects the acceptor concentration and therefore the energy transfer efficiency.

For the fluorescence experiments stock solutions of protein mutants and wild-type protein solubilized in cholate buffer were mixed with solutions of lipids in the same buffer, as described previously (Spruijt et al., 1989). Repeated dialysis of the mixtures in cholate-free buffer was performed to remove the cholate in the sample. The lipid loss during dialysis can vary near 20% (Spruijt et al., 1989), and this fact should be taken into account during the analysis of the experimental data.

Table 2.1. Sample composition of M13 major coat protein incorporated into DOPC:DOPG bilayers given in terms of r_{LP} and r_{ul}, and observed energy transfer efficiencies E for mutants with acceptor positions n_A at 24, 38 and 46. For mutant G38C two FRET titration experiments were carried out at different values of r_{LP} and r_{ul}.

Data set	❶	❷	❸	❹
n_A	24	38	38	46
r_{LP}	3600	209	3213	105
r_{ul}	0.6	6	1	1.3
E	0.558	0.121	0.254	0.152
r_{LP}	1059	128	553	80
r_{ul}	4.5	10	10	2.2
E	0.165	0.094	0.056	0.147
r_{LP}	621	71	303	55
r_{ul}	8.4	19	18	3.9
E	0.099	0.071	0.043	0.135
r_{LP}	340	45	159	38
r_{ul}	16	33	36	6
E	0.058	0.056	0.027	0.127
r_{LP}	179	28	65	25
r_{ul}	32	54	88	10.4
E	0.033	0.047	0.020	0.116

2.2.2. FRET experiments

Optical spectroscopy. Absorption spectra were recorded on a Varian Cary 5E UV-Vis-NIR spectrophotometer and fluorescence emission and fluorescence excitation measurements were performed on a Fluorolog 3.22 manufactured by Jobin Yvon-Spex as

described elsewhere (Gustiananda et al., 2004; Vos et al., 2005). For fluorescence excitation measurements the detection wavelength was set at the maximum of the acceptor (AEDANS) fluorescence of a particular mutant and the excitation wavelength was scanned from 260 to 400 nm. The detection wavelength was different for each mutant, because the AEDANS fluorescence maximum varies with bilayer depth (i.e. local polarity) of the AEDANS label and therefore with the residue number of the labeled cysteine. The AEDANS fluorescence for mutants 24 and 46 was observed at 490 nm; for mutant 38 this was 470 nm. The applied slit widths of the detection and excitation monochromators corresponded to 5 and 2 nm band pass, respectively. The spectra were automatically corrected on the Fluorolog 3.22 for variations in the lamp output by dividing the sample signal by that of an internal reference detection system. All excitation spectra were corrected for background fluorescence using an equimolar solution of pure wild-type protein (no AEDANS present). Moreover, tryptophan fluorescence is neglectable at the detection wavelength (see Fig. 2.4 A), therefore the observed radiation exclusively belongs to AEDANS. The temperature during all measurements was 20°C. Because of the small protein concentrations used in our experiments (about 1 μM), errors caused by the inner filter effects can be neglected.

Analysis of AEDANS excitation spectra. The derivation of the mathematical expressions for the analysis of the experimental excitation spectra is given in Appendix A (section 2.6). For our analysis we used the energy transfer efficiency E, which can be calculated from the fluorescence intensities (Lakey et al., 1993) by

$$E = \frac{1}{1 + r_{ul}} \left(\frac{F^{290}}{F^{340}} - \frac{\varepsilon_A^{290}}{\varepsilon_A^{340}} \right) \frac{\varepsilon_A^{340}}{\varepsilon_D^{290}}, \tag{2.1}$$

where r_{ul} is the ratio of the number of unlabeled to labeled proteins. For every sample the ratio of the fluorescence intensity at 290 nm, F^{290}, (mainly donor excitation) to that at 340 nm, F^{340}, (exclusively acceptor excitation) was calculated, being a measure of the donor-to-acceptor energy transfer. The ratio F^{290}/F^{340} was corrected for direct excitation of AEDANS at 290 nm by subtracting the ratio of the extinction coefficients $\varepsilon_A{}^{290}/\varepsilon_A{}^{340}$ = 0.20 (this ratio was calculated using mutant Y21A/Y24A/W26A/G23C). Finally, the ratio of the extinction coefficients of the acceptor at 340 nm ($\varepsilon_A{}^{340}$) and donor at 290 nm ($\varepsilon_D{}^{290}$) have to be taken into account in Eq. 2.1 ($\varepsilon_A{}^{340}/\varepsilon_D{}^{290} = 1.2$).

2.3. Methodology

2.3.1. Model for the transmembrane domain of M13 coat protein incorporated into a lipid bilayer

The proposed simplified structural model for the transmembrane domain of M13 coat protein consists of an ideal α-helix (Fig. 2.1) (Glaubitz et al., 2000; Koehorst et al., 2004; Marassi and Opella, 2003; Stopar et al., 2003). The complete set of structural parameters that determines the protein-lipid system is presented in Table 2.2. In the protein model, we assume two specific sites: a donor and an acceptor site that will enable us to calculate the theoretical energy transfer and relate that to the FRET experiments. For M13 coat protein, which consists of 50 amino acid residues, the donor is the Trp-26 and the acceptor is introduced at an arbitrary position in the transmembrane protein domain via cysteine mutagenesis and labeling with a fluorescent label (in our case: AEDANS). Acceptor sites are empty for non-labeled or wild-type proteins.

As a model for proteins incorporated into a lipid bilayer, a square region of a bilayer containing a certain number of proteins (N_P) is considered. By using a three-dimensional mathematical description, protein molecules as shown in Fig. 2.1 are inserted randomly (both in location as well as in orientation) into the lipid bilayer in the way that the angle θ between the membrane normal and their main axis O of the transmembrane domain is between 0 and 90º. The direction of the protein tilt is given by ψ. A value $\psi = 0$ means that protein is tilted towards the C_α of the reference (n_0) amino acid residue. The depth of protein insertion is given by parameter d. It is assumed, that when inserted into the membrane, the proteins occupy a cylindrical region in both bilayer leaflets with a protein exclusion distance D_P. Within this region no lipids or other proteins can be located.

In the protein-lipid model the direction of the tilt and the orientation of the N-terminal domain of each protein in the coordinate system of the bilayer are set randomly. Two algorithms of protein insertion were considered. In the first one three reference points located in the transmembrane domain were selected, and during insertion the distances between these reference points of the inserted protein and similar points on the nearest proteins were compared with D_P to determine the overlapping situation. In case of a clash, the algorithm selected a different protein direction or, if still unsuccessful after a number of tries, a new protein position. In the second algorithm the proteins were simply inserted randomly at distances larger than D_P. In this case their tilted transmembrane domains could in principle overlap. This first algorithm turned out to be quite time consuming (from 2 to 20 times,

depending on the L/P ratio) without significant changes in the energy transfer results (less than 0.02 for the extreme case of L/P 25). Therefore, we decided to use the simplified algorithm in all further fitting procedures.

The area of the considered square region of the membrane is calculated from the experimental L/P ratio (r_{LP}), the protein exclusion distance (D_P), the area per two lipid molecules (S_L) and the ratio of lipids lost during dialysis to their initial quantity (i.e. the lipid loss L) in the following way:

$$S = N_P\left(S_L r_{LP}(1-L)/2 + \pi D_P^2/4\right). \tag{2.2}$$

Furthermore, to be able to work with mixtures of labeled and unlabelled protein molecules, the ratio r_{ul} between the number of unlabeled and labeled proteins needed to be introduced into the model.

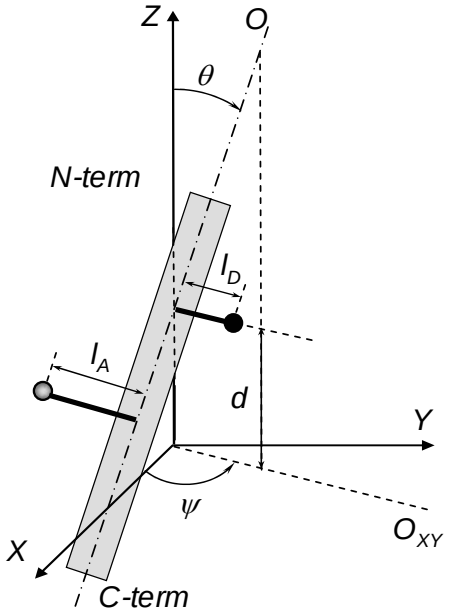


Figure 2.1. (A) Schematic drawing of the transmembrane domain of M13 major coat protein consisting of an ideal α-helix (Glaubitz et al., 2000; Koehorst et al., 2004; Marassi and Opella, 2003; Stopar et al., 2003). As an example, the donor (Trp-26, black circle, located on the N-terminal side at a distance l_D from the protein helix axis) and acceptor (AEDANS, grey circle, located on the C-terminal side at a distance l_A from the protein helix axis) are attached at positions 26 and 38, respectively. The membrane axis system is indicated by *X*, *Y*, and *Z*. The *XY* plane at $Z = 0$ corresponds to the centre of the lipid bilayer in which the protein is inserted. Parameter d is the distance from the origin of the coordinate system of the protein to the centre of the lipid bilayer. Axis *O* is the helix axis of the protein domain, and θ is the tilt angle, *i.e.* the angle between the helix axis and the normal to the membrane. O_{xy} is the projection of the helix axis on the *XY* plane. Angle ψ is the protein tilt direction, *i.e.* the direction of the tilting of the helix. The complete set of structural parameters that determines the protein-lipid system is presented in Table 2.2.

Table 2.2. Definition of the parameters used in the model for the protein-lipid system. In the simulations parameters θ, d, ψ, L, and k are varied. Parameters n_A, r_{LP} and r_{ul} are determined by the experiment; the other parameters are fixed as shown in the table.

Para-meter	Range or value	Unit	Description
n_0	26	–	The position of a reference amino acid residue. The projection of its C_α to the helix axis of the protein O gives the origin of the coordinate system of the protein. Position n_0 = 26 was selected for the transmembrane domain of M13 major coat protein.
h	1.5	Å	Translation per amino acid residue along the helix; this is 1.5 Å for a perfect α-helix.
n_r	3.6	–	Number of amino acid residues per one turn; this is 3.6 for a perfect α-helix.
n_D	26	–	Donor position; position of amino acid residue given by the donor. For M13 coat protein the donor is Trp-26, which is located in the transmembrane domain.
n_A	1 – 50	–	Acceptor position; position of amino acid residue labeled by the acceptor. For the transmembrane domain of M13 coat protein the acceptor positions are 24, 38 and 46.
l_D	6.5	Å	Donor arm, the average distance from the donor moiety to the helix axis. A value l_D = 6.5 Å was taken (Koehorst et al., 2004).
l_A	9.5	Å	Acceptor arm, the average distance from the acceptor moiety to the helix axis. A value l_A = 9.5 Å was taken (Koehorst et al., 2004).
θ	0 – 90	°	Protein tilt angle; the angle between the helix axis and the normal to the membrane.
d	0 – 30	Å	Distance from the origin of the coordinate system of the protein to the centre of the bilayer.
ψ	-180 – 180	°	Protein tilt direction; the direction of the protein transmembrane domain tilting. A value ψ = 0 means that protein is tilted towards the C_α of the reference (n_0) amino acid residue.
N_P	500	–	Number of proteins in the system. All simulations were performed for models containing 500 proteins.
S_L	72	$Å^2$	Area occupied by a lipid in one leaflet of a bilayer; the average area for the DOPC:DOPG system is 72 $Å^2$ (Fernandes et al., 2003).
L	0.0 – 1.0	–	Lipid loss; ratio of lipids lost during dialysis to their initial quantity.
D_P	10	Å	Protein exclusion distance; minimal protein-protein distance. For M13 coat protein a value D_P = 10 Å was taken.
r_{LP}	≥ 0	–	Lipid to protein ratio.
r_{ul}	≥ 0	–	Ratio between the number of unlabeled and labeled proteins.
k	0.0 – 1.0	–	Protein-protein association probability, defined as the percentage of clustered proteins with respect to the total number of proteins (see Fig. 2.2).
R_0	24	Å	Förster distance. A value of 24 Å is calculated using the data about the photophysical properties of the donor and acceptor.

Similar to the experimental reconstituted protein-lipid system, protein molecules can be inserted into the model membrane randomly with "parallel" and "anti-parallel" orientations; this means that the N-terminal domain of the protein can be located either in the upper or in the lower leaflet of the membrane with equal probabilities. The result of these equiprobable orientations is that the membrane system contains two layers of donors and two layers of acceptors.

A protein-protein association probability k is introduced to take into account the ability of the membrane proteins to form oligomers or clusters. The algorithm for this association is as follows. All proteins are divided into two groups: free and associated. Initially, the coordinates of the free proteins in the XY plane of the membrane are randomly generated. Before incorporation of a new protein into the membrane model, it is checked whether the position for the protein is free (all previously incorporated proteins are not closer than D_P). If the position is occupied, random coordinates are selected again. For associated proteins the algorithm is slightly changed: the XY coordinates are selected in a way, to incorporate the protein at a distance D_P next to one of previously incorporated proteins. The value of k ranges from 0 to 1, indicating no association and complete association (all proteins are clustered together), respectively. The effect of protein association is exemplified in Fig. 2.2.

Apart from the structural parameters and parameters related to the composition of the protein-lipid system, one additional physical parameter needs to be introduced: this is the Förster distance R_0 of the donor-acceptor pair. Its physical meaning is discussed below.

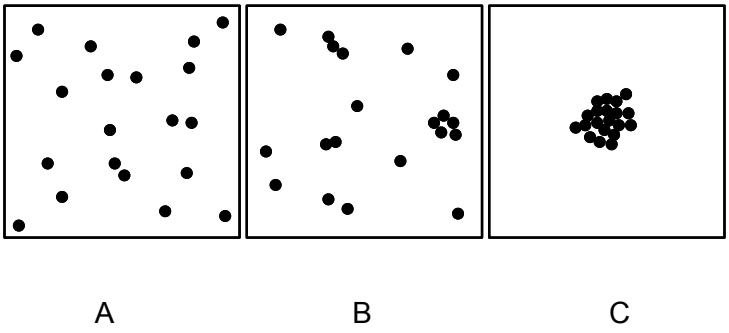


Figure 2.2. Schematic illustration of the effect of protein association resulting from the model described in the text. (A) Random distribution of proteins with k = 0, (B) partially associated proteins with k = 0.5, (C) completely associated proteins (k = 1). Proteins are schematically indicated by solid dots. The figures show that at increasing values of k the proteins aggregate into clusters in a non-specific way.

2.3.2. Models for FRET

Basic model for energy transfer. Being in an excited state a fluorescent molecule has a dipole-dipole interaction with other molecules in close proximity, which can lead to energy transfer from the excited molecule to the non-excited ones. If we assume that the emission spectrum of the donor overlaps with the absorption spectrum of the acceptor, the photon absorbed by the donor can be transferred to the acceptor with a rate constant k_{ET} depending on the sixth power of the distance between the donor and acceptor

$$k_{ET} = \frac{1}{\tau_D}\left(\frac{R_0}{R}\right)^6, \tag{2.3}$$

where τ_D is the lifetime of an isolated donor, R the distance between the donor and acceptor. The so-called Förster distance R_0 is given by

$$R_0 = 9780\left(\kappa^2 n^{-4} Q_D J\right)^{1/6}. \tag{2.4}$$

In this equation κ^2 is the orientation factor describing the relative orientation of the transition dipole moments of the donor and the acceptor, n is the refractive index of the environment, Q_D is the quantum yield of an isolated donor, and J is the integral expressing the degree of donor emission and acceptor absorption spectral overlap (Lakowicz, 1999).

Consider now a system of multiple donors and acceptors that are fixed at their positions. Let us number the donors i=1..N_D, and acceptors j=1..N_A. Here N_D is the number of donor molecules, and N_A the number of acceptor molecules. The probability for each donor to transfer energy to one of the acceptors can then be calculated as follows:

$$p_i = \frac{\sum_{j=1}^{N_A} k_{i,j}}{\frac{1}{\tau_D} + \sum_{j=1}^{N_A} k_{i,j}} = \frac{\sum_{j=1}^{N_A} \left(R_0 / R_{i,j}\right)^6}{1 + \sum_{j=1}^{N_A} \left(R_0 / R_{i,j}\right)^6}, \tag{2.5}$$

where $R_{i,j}$ is the distance between the i-th donor and j-th acceptor.

The mean probability of energy transfer events for all donor molecules gives the energy transfer efficiency E for the entire system:

$$E = < p_i >_{N_D}. \tag{2.6}$$

Steady-state FRET simulation. To analyze the experimental steady-state fluorescence data, steady-state FRET simulation is employed. The main advantage of this approach over Monte Carlo time-resolved simulation is its simplicity and high speed. The simulation starts with the generation of the structural model for the protein-lipid system. This model provides the coordinates of each donor and acceptor. The energy transfer efficiency E is then calculated using Eqs. 5 and 6. Because of the stochastic nature of the structural model, the resulting energy transfer efficiency contains stochastic deviations. Therefore the simulations are executed several times to make the results statistically relevant. The flow diagram of the simulation is shown in Fig. 2.3 and described below.

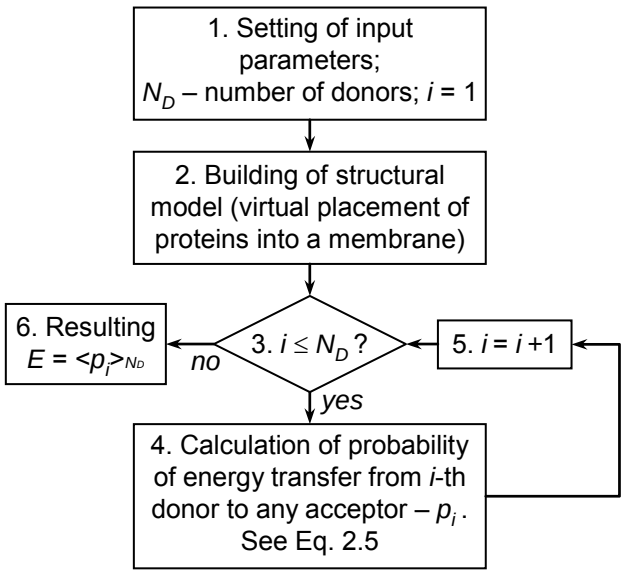


Figure 2.3. Flow diagram of a single simulation of energy transfer in a protein-lipid system.

1. The parameters of the system are set (block 1).
2. The structural model of a membrane with embedded proteins is created in accordance with the input parameters. The coordinates and orientation of the proteins provide information about the locations of donors and acceptors in the system (block 2).
3. For each donor (denoted as i) the distances to all acceptors are considered and the probability of energy transfer (to any of them) is calculated using Eq. 2.5 (blocks 3-5).
4. The mean probability of energy transfer among all donors results in the energy transfer efficiency for the whole system (Eq. 2.6).

5. Steps 2-4 (and blocks 2-6 in the flow diagram) are repeated for a number of times to decrease the effect of the randomness of the protein distribution. In our calculations we executed the simulation for 100 times.

Analytical model of FRET in planar systems. An analytical expression for FRET in a planar system was initially developed by Wolber and Hudson (Wolber and Hudson, 1979) and further enhanced by Davenport et al. (Davenport et al., 1985). In these models acceptors were considered as molecular systems of infinitesimal size uniformly distributed in a plane. The original equations by Davenport et al. can be modified to describe the energy transfer in the systems of M13 coat protein incorporated into lipid bilayers. The resulting analytical expression for the energy transfer efficiency E in the considered system is

$$E = 1 - \frac{1}{\tau_D} \int_0^\infty \rho_D(t) \times q_\sigma(t) \times \frac{r_{ul} + q_{intra}(t)}{1 + r_{ul}} dt \,, \tag{2.7}$$

where ρ_D is the fluorescence decay of a single donor, τ_D is the donor lifetime, and q_σ and q_{intra} are the quenching contributions of inter and intramolecular energy transfer, respectively. The derivation of Eq. 2.7 and a further description of the expressions for ρ_D, q_σ, and q_{intra} are given in Appendix B (section 2.7).

2.3.3. Simulation-based fitting approach to experimental data analysis

The FRET model developed for M13 coat protein incorporated into lipid bilayers is used to analyze experimental data via the SBF approach. The scheme of SBF has been discussed in detail recently (Nazarov et al., 2004). As a measure of the goodness of the fit the following criterion was introduced:

$$\chi^2 = \sum_{i=1}^{N} \left(E_i^e - E_i^s \right)^2 , \tag{2.8}$$

where N is the number of data points, E_i^e the experimentally obtained energy transfer efficiency, and E_i^s the simulated energy transfer efficiency. To fit the modeled energy transfer efficiencies to the experimental ones, an optimization algorithm should be used. In our case gradient optimization techniques are not applicable to fit the data, because of the stochastic behavior of the error function χ^2. Therefore, to perform a simultaneous fit of all experimental data the Nelder-Mead "simplex" method (Nelder and Mead, 1965) is used. This method provides a reasonable convergence and is not extremely time consuming. To increase

the robustness of the method and the precision of the solution a global analysis approach is chosen, and therefore all experimental data were fitted simultaneously (Beechem and Brand, 1986).

Because of the stochastic behavior of the FRET model, the error function χ^2 is stochastic as well, and the parameters obtained after each fit contain random deviations that are dependent on the sensitivity of the energy transfer to variations of the parameters. Therefore, to deal with this stochastic effect and to avoid possible local minima, the fitting procedure is performed 100 times with different initial estimations of the fitting parameters. The methodology used for the analysis of the resulting solutions and the selection of the representative solutions in terms of an optimal 20% "elite" subset is given in Appendix C (section 2.8).

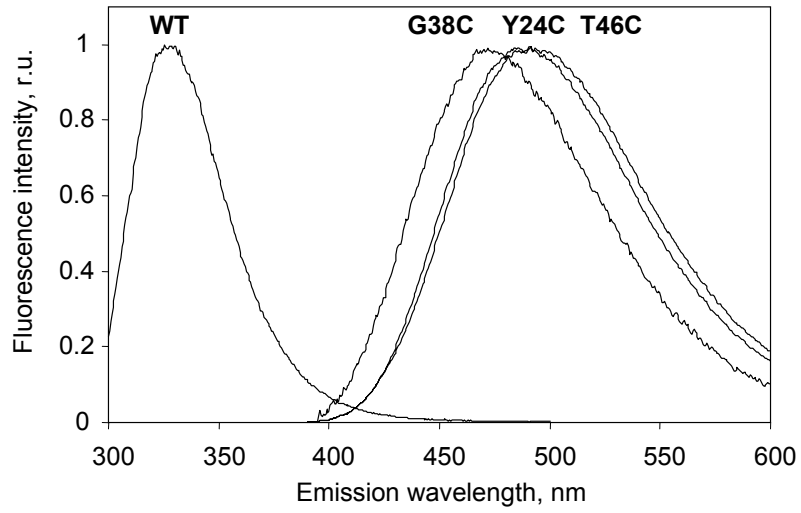
All models were realized as C++ classes. The Borland C++ Builder 6.0 environment was used to combine the developed models, OpenGL visualization and SBF fitting algorithms into a software tool called FRETsim. The C++ classes and software are available from the authors upon request.

2.4. Results

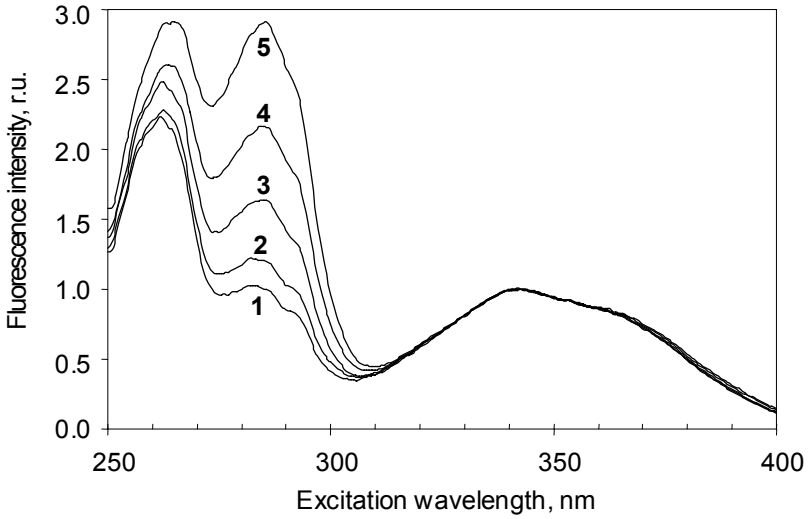
2.4.1. Experimental energy transfer efficiencies

An example of the experimentally obtained excitation spectra at different L/P ratios is presented in Fig. 2.4. The increase of the fluorescence intensity at the donor absorption wavelength (290 nm) clearly shows the increasing effect of energy transfer.

The mutants that were selected for our experiments (Y24C, G38C, and T46C) have their cysteines, and therefore the AEDANS labels, on the boundaries or close to the center of the transmembrane α-helix, which ranges from about amino residue 25 to 45 (Koehorst et al., 2004; Papavoine et al., 1997; Vos et al., 2005). For mutant G38C two FRET titration experiments were performed at different values of r_{LP} (and also acceptor concentrations) to study its effect of protein association, given by parameter k. As a result of titration experiments on Y24C, T46C, and the double experiment on G38C mutants four data series were obtained. The experimental L/P ratios r_{LP}, the unlabelled-to-labeled protein ratios r_{ul}, and resulting energy transfer efficiencies are presented in Table 2.1. The behavior of the energy transfer efficiency for different mutants as a function of r_{ul} is illustrated in Fig. 2.5.



A



B

Figure 2.4. (A) Emission spectrum of wild type proteins (WT) showing the Trp fluorescence, and emission spectra of mutant proteins Y24C, G38C, and T46C with AEDANS-labeled Cys at positions 24, 38, 46 after subtraction of the fluorescence of equimolar WT samples. Note that almost no Trp fluorescence can be observed at the AEDANS emission maxima. (B) Experimental excitation spectra obtained for mutant 38 at different titration points of wild-type proteins. The emission was detected at 470 nm. The labels 1 to 5 correspond to r_{ul} values of 6, 10, 19, 33, and 54, respectively. The lipid-to-protein ratios r_{LP} are 209, 128, 71, 45, and 28, respectively (see data set ❷ in Table 2.1). The sample showing the highest peak at 290 nm (spectrum 5) has the highest protein density (lower r_{LP}) and r_{ul}. Although the efficiency of energy transfer (Fig. 2.5) for this case is smallest, the overall energy absorbed by the donors in such a system, and therefore the transferred (intermolecular), is higher than for the other values of r_{LP} and r_{ul}.

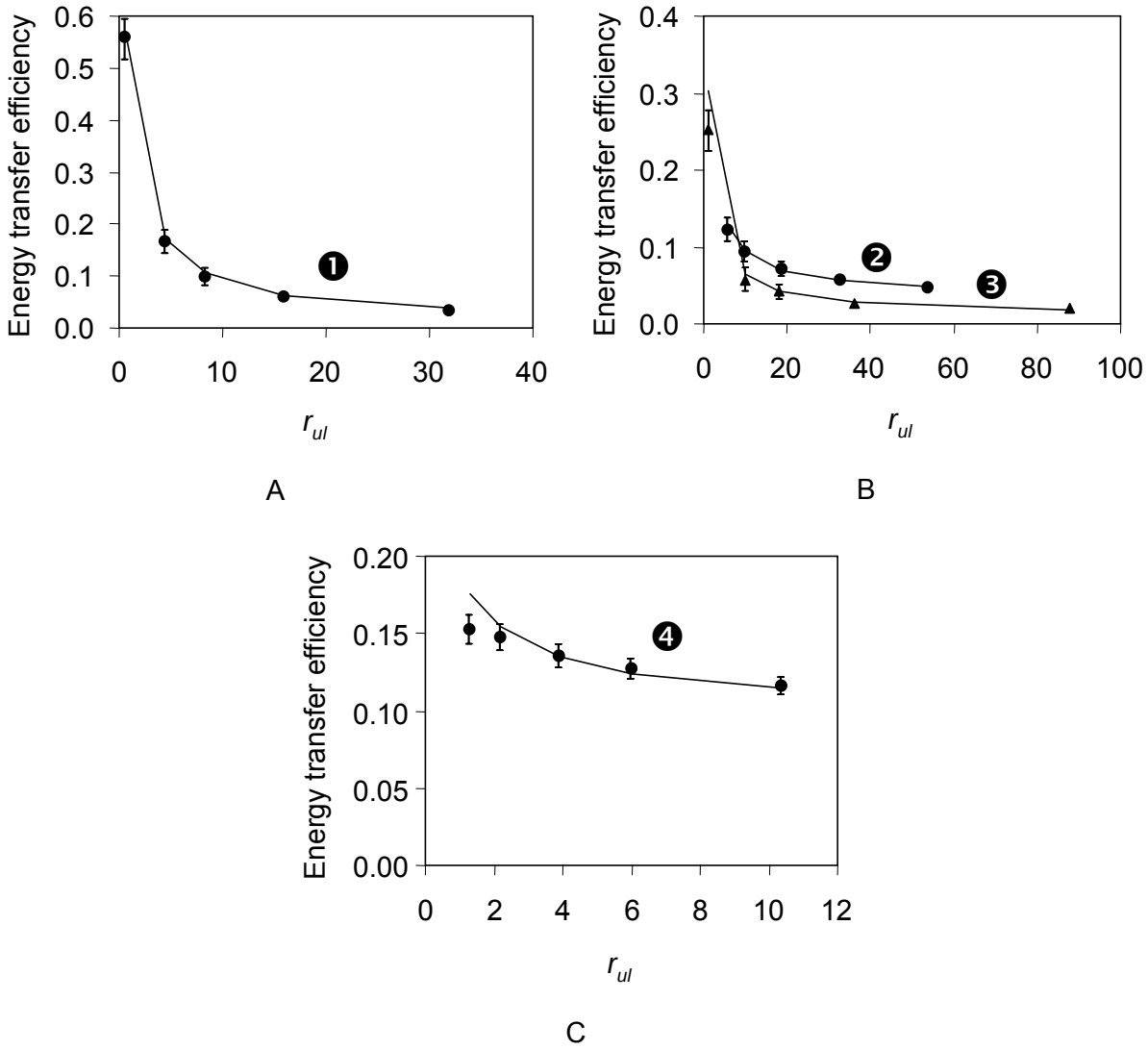


Figure 2.5. Experimental energy transfer efficiencies E (filled dots and triangles) and their approximation by the model (solid line) after global analysis versus the ratio between unlabeled and labeled proteins r_{ul}. (A) mutant 24, (B) mutant 38, (C) mutant 46. The labels ❶–❹ refer to the corresponding data sets in Table 2.1. In (B) the dots indicate data set ❷ and the triangles data set ❸. The error bars correspond to the maximal deviations of the data points observed during the experiments.

2.4.2. Förster distance

The value of Förster distance R_0, needed for simulation of energy transfer, was calculated using Eq. 2.4. In this equation $Q_D = 0.23$ was taken, which is the quantum yield of tryptophan in dimyristoyl phosphatidylcholine (DMPC) bilayers (Fisher and Ryan, 1999). The overlap integral J is calculated from the emission spectrum of the wild-type protein and the absorption spectrum of the AEDANS-labeled Y21A/Y24A/W26A/G23C mutant, which has no tryptophan at position 26. This results in a value of 5.96×10^{-15} $M^{-1}cm^3$. For small proteins and peptides, as is the case for M13 coat protein, the orientation factor κ^2 can be approximated

by its isotropic dynamic average, giving a value of 2/3 (Gustiananda et al., 2004; Kamal and Behere, 2002; Lakshmikanth et al., 2001; Loura et al., 1996; Vos et al., 2005). For simplicity the refractive index of the medium is assumed to be constant, and equal to 1.4 (Davenport et al., 1985; Lakowicz, 1999). These parameters result in a Förster radius R_0 of 24 Å. It should be noted, that the excitation band of AEDANS with its maximum around 340 nm does not change with the position of the labeled cysteine. This implies that the Förster distance for the donor-acceptor pair is equal for all mutants.

2.4.3. Determination of bilayer topology of the protein

All four sets of experimental data were fitted simultaneously. The fitting procedure included 50 iterations of the Nelder-Mead "simplex" method. To avoid local minima the fitting procedure was independently repeated 100 times with different initial estimations of the desired parameters: L, k, θ, ψ, and d. The values of initial estimations were randomly selected from the parameter ranges, presented in Table 2.2. The calculation of each single solution took approximately 20 min on a computer with a Pentium 4 processor (each simulation takes 1-5 s). Because the calculation of each solution is an independent task, the fitting was parallelized between several computers. The solutions found were analyzed as described in paragraph 3.3. The resulting χ^2 for the "elite" set varies from 0.0039 to 0.0048, the discarded solutions had a χ^2 ranging from 0.0048 to 0.1.

The resulting values together with the standard deviations inside the "elite" set of solutions are presented in Table 2.3. This table also shows a compilation of the values known from the literature. The best fitting results are presented in Fig. 2.5 together with the experimental data.

Table 2.3. Resulting parameters of the model for the protein-lipid system applied to the transmembrane domain of M13 major coat protein incorporated into DOPC:DOPG bilayers and the corresponding values known from the literature.

Parameter	Value found	Previously reported value	Reference
L	0.28 ± 0.03	~ 0.2	(Spruijt et al., 1989)
k	0.03 ± 0.01	~ 0	(Fernandes et al., 2004)
θ	18 ± 2°	19 ± 1°	(Koehorst et al., 2004)
		26°	(Marassi and Opella, 2003)
		20 ± 10°	(Glaubitz et al., 2000)
ψ	61 ± 7°	60°	(Koehorst et al., 2004)
d	8.5 ± 0.5 Å	8.9 Å	(Koehorst et al., 2004)

2.5. Discussion

2.5.1. Measuring strategy

In this study we aimed at the development of a methodology based on a combination of FRET spectroscopy and computer simulation, thereby providing information about the position and protein-protein associations in a membrane system. By assuming a helical structure for the fluorescent-labeled protein (or its domain) the proposed approach is able to determine both its topology and bilayer embedment in terms of protein tilt angle, direction of tilt and protein depth in the membrane. Moreover, the method provides a quantitative analysis of the protein-protein associations, which can hardly be performed by other spectroscopic methods. In the case of a non-dilute protein-lipid system with randomly distributed proteins the energy of donor excitation can be transferred both intra and intermolecularly. Because the aggregation behavior of M13 coat protein in lipid vesicles is not well documented, and cannot be excluded even at high L/P ratios, the efficiency of the intermolecular energy transfer component may partly arise from relatively short donor-to-acceptor distances in protein aggregates.

Being incorporated into the membrane, the proteins form two planes of donor and two planes of acceptor molecules, originating from "parallel" and "anti-parallel" orientations of the proteins. The intermolecular energy transfer is influenced, among other factors, by the distances between the donor and acceptor planes, which are determined by the z-coordinates of the fluorescent labels. Structural parameters describing the embedment and orientation of the protein, such as d, θ, and ψ (see Fig. 2.1) can change the positions of the planes, and therefore can be tracked by analyzing energy transfer processes.

The selection of mutants Y24C, G38C, and T46C was given by two rationales. First, the selected labeling sites should be located in an α-helical part of the M13 coat protein. This condition arises from the assumption of an α-helical protein model. Second, the selected sites should present maximally diverse intramolecular distances and acceptor positions inside the membrane, to increase the precision of the parameter determination. Therefore, sites should be located preferably at the edges of such a helical part. An α-helical conformation was suggested for positions from about 25 to 45 in the transmembrane domain (Koehorst et al., 2004; Papavoine et al., 1997; Vos et al., 2005). Therefore we selected the mutants Y24C, G38C, and T46C as labeling sites. To study possible effects of protein aggregation additional experiments were performed for the G38C mutant at high and low L/P ratios (see Table 2.1).

For an ideal case of independent parameters and independent experiments without any distortion in the obtained data, the number of experiments N should be equal to n ($N = n$), where n is the number of unknown parameters. However if the data set contains noise, the number of equations should be larger than the number of parameters (*i.e.* $N > n$). Obviously, the more data provided, the higher the precision one would get. In our specific situation each of the data series (as shown in Table 2.1 and Fig. 2.5) can be considered as two independent points representing the intra and inter-molecular energy transfer. Therefore, for the situation of an α-helical protein model with five unknown structural parameters (giving $n = 5$, *i.e.* θ, d, ψ, L, and k), at least one independent data series coming from each of the three selected mutants is needed (giving $N = 6$). Of course, including additional series would enhance the precision of the determination of the parameters. Thus as a rule of thumb at least three donor-acceptor pairs would be needed that are regularly spread over the protein transmembrane domain.

To determine the energy transfer parameters, fluorescence excitation spectroscopy was used by monitoring the acceptor excitation at wavelengths 470-490 nm. Here the acceptor fluorescence was monitored, thereby optically selecting only the acceptor-labeled mutants, and discriminating between fluorescence resulting from donor-to-acceptor energy transfer and fluorescence resulting from direct excitation. There are two advantages of recording the acceptor fluorescence excitation over the donor fluorescence. First, the intensity of the background fluorescence between 450 and 550 nm is less than in the UV region (tryptophan/donor fluorescence is between 300 and 350 nm). Second, for an experiment in which the donor is monitored, varying the wild-type protein concentration or varying the mutant protein concentration would change the concentration of that donor, while in our approach the concentration of the monitored acceptors is kept constant. The small lifetime of tryptophan (~3.6 ns) allows us to assume that there is no lateral mobility in the system that can significantly change the donor-acceptor distance.

To separate intra and intermolecular energy transfer contributions, we performed titration experiments in which mixtures of a fixed amount of labeled protein mutants and different amounts of wild type protein were re-constituted into lipid vesicles. Both unlabeled mutant and wild-type protein can be considered spectroscopically identical as donor-containing molecules without acceptor label, however, labeled mutants contain both a donor and acceptor.

2.5.2. Validation of the simulation model

Before applying the protein-lipid model and the SBF approach to real experimental data, both the model and the approach should be validated. As a first step, the energy transfer efficiency is calculated for a system with different L/P ratios r_{LP} (for simplicity we consider a constant $r_{ul} = 0$) and compared with results of the modified Davenport's analytical model, Eq. 2.7. The comparison is carried out for different values of r_{LP}, which influence the acceptor surface density. The resulting energy transfer efficiencies are plotted in Fig. 2.6, using a value for D_P, and consequently the exclusion distance in Davenport's model, of 10 Å, which is about the diameter of a transmembrane protein domain. The plot shows a deviation of the analytically obtained energy transfer efficiencies from the simulated ones. This finding provoked us to perform an additional study on the applicability of the analytical solution. As was mentioned before, the analytical solution is based on a number of simplifications; one of those is the assumption of an infinitely small acceptor size. To check this situation a comparison is carried out by assuming as small transmembrane protein domain with an exclusion distance D_P of 1 Å. For such a system a complete correspondence between the simulated and analytically calculated energy transfer efficiency is observed (Fig. 2.6).

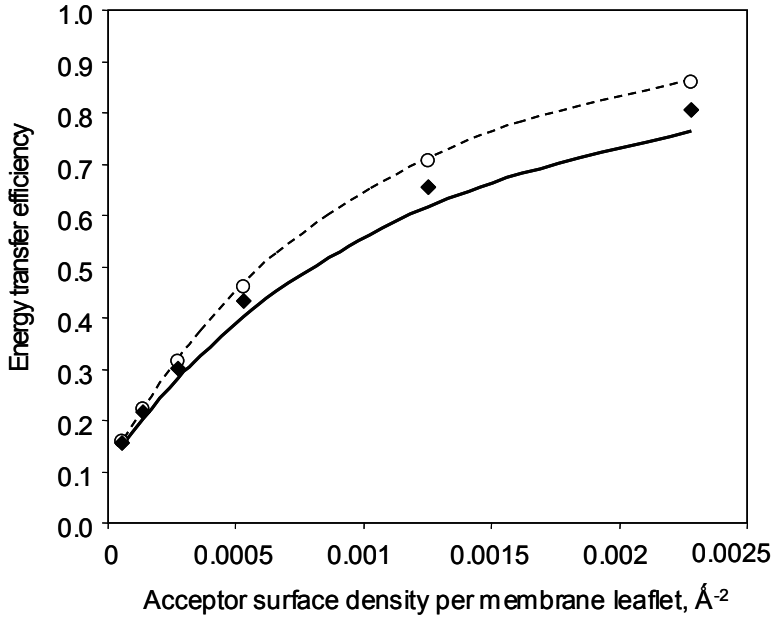


Figure 2.6. Comparison of simulation results with analytical solutions for different sizes of proteins. Solid line: analytical result for D_P = 10 Å; ♦: simulation results with the same protein exclusion distance D_P = 10 Å; dotted line: analytical energy transfer efficiency; o: simulated energy transfer efficiency for D_P = 1 Å. All calculations were performed with the following protein parameters: n_A = 46, n_D = 26, θ = 16°, ψ = 50°, d = 10 Å, S_L = 72 Å^2, and R_0 = 24 Å. The corresponding parameters for the analytical model are: h_{I} = 26.8 Å, h_{II} = 13.1 Å, and R_{intra} = 33.9 Å.

From the comparison it is clear that the steady-state simulation model of FRET gives the same results as the extended well-known analytical solution of Davenport et al. (Davenport et al., 1985) (Eq. 2.7) in the case of small acceptor-labeled molecules. However, if the size of the molecules becomes comparable to the Förster distance, the simulation-based approach should be used rather than the analytical model. It is clear that the limiting situation for small molecular sizes of the simulation-based approach corresponds to the analytical solution. The simulation-based approach is more general and powerful than the analytical model and can be applied for the analysis of donor-acceptor systems with any geometry.

2.5.3. Testing of the simulation-based fitting approach

A numerical test was performed to prove the applicability of the SBF approach to the problem of M13 coat protein structure determination and to find the optimal "elite" subset size. In this test synthetic FRET data were generated using the model with values of parameters close to those determined experimentally for M13 coat protein. The simulation was performed for 1000 proteins and the results were averaged for 1000 simulations. This provided us with synthetic data containing a very small randomness. Then these synthetic data were analyzed via the SBF approach as mentioned before, and the solutions were handled as shown in Appendix C (section 2.8). The smallest deviation from the original values of the parameters was found for a 20% "elite" subset. The results are presented in Table 2.4. The random spread of the solutions inside the "elite" set was close to that obtained during the analysis of the experimental data (Table 2.3).

Table 2.4. Original and calculated values of the model parameters after analysis of synthetic FRET data by means of a SBF approach. To introduce noise in the synthetic data, a standard deviation of ~10% is used for points with a low r_{ul} and ~5% for points with a high r_{ul} (see the error bars in Fig. 2.5).

Para-meter	Original value in synthetic data simulation	Value found after SBF analysis with no noise added to synthetic data	Value found after SBF analysis with additional noise in synthetic data
L	0.3	0.29 ± 0.01	0.30 ± 0.03
k	0.05	0.05 ± 0.01	0.07 ± 0.02
θ	20°	19 ± 3°	16 ± 3°
ψ	60°	61 ± 8°	47 ± 15°
d	9 Å	8.9 ± 0.2 Å	9.2 ± 1 Å

The values found for the association coefficient k and lipid loss L (see Table 2.4) are very close to the original ones, and have a relatively small error. These two parameters influence the surface density of the label and therefore have a strong effect on the intermolecular FRET. Despite some correlation between k and L, the method is able to determine both parameters quite well. From the results in Table 2.4 it follows that the values of the protein depth d, protein tilt angle θ are close to the original ones. The direction of protein tilting ψ has a substantial large variation. The reason for the spread in ψ is that this parameter does not significantly influence the position of the donor and acceptor planes.

To study the possible effect of experimental noise on the resulting data, we introduced Gaussian noise to the synthetic data, and performed a number of fittings. Each SBF was performed with its own random deviations in the data. The standard deviation of each data point was calculated in according to maximal deviations observed in the FRET experiments, which are ~10% for points with a low r_{ul} and ~5% for points with a high r_{ul} (see the error bars in Fig. 2.5). The results of the fitting of the noisy data are given Table 2.4. As can been seen, the parameters L, k and d are only slightly affected by introducing noise, indicating that they can be determined quite precisely from the FRET experiments. This stability to noise can be explained by the fact that we use a global analysis approach and that for each mutant we have 5 data points. The angular parameters (θ and ψ) tend to deviate from the original value, indicating that they are relatively more sensitive to noise in the experimental data.

From this test it can be concluded that the application of the described biophysical model together with the SBF approach to data analysis is capable to determine the protein location in a bilayer, and the protein-protein association. This result gives us confidence to apply the methodology to analyze our experimental FRET data.

2.5.4. Parameters determined

Table 2.3 summarizes the resulting parameters and the corresponding values known from the literature. Variation of the parameters within the error limits given in Table 2.3 does not result in values of χ^2 higher than 0.0048 (in fact, all acceptable solutions have χ^2 values between 0.0039 and 0.0048, see Fig. 2.7 A). For example, increasing θ by 20° to 38° increases χ^2 to 0.0113. This χ^2 value is far above the limit of 0.0048 that was taken as acceptable.

The actual value of the parameter describing the lipid loss during dialysis L is unknown and has to be determined using the SBF approach from our experimental data. The value found is 0.28, which means that approximately 28% of the lipids are washed-out from

the sample into the buffer during dialysis. This value is in reasonable agreement with the lipid loss of 20% as estimated from biochemical analysis (Spruijt et al., 1989). All our experiments were performed under identical conditions and using the same protocol. This allows us to assume that the lipid loss is constant for all experiments. The small value of the association constant k indicates that the proteins have no tendency to aggregate under the experimental conditions. This is in agreement with earlier observations of the protein in DOPC:DOPG mixtures (Fernandes et al., 2004). Again some correlation between parameters k and L was found. This effect is included in the uncertainty limits for the parameters in Table 2.3.

The resulting protein depth d of 8.5 Å is very close to the value of 8.9 Å as found from fluorescent experiments in DOPC:DOPG (Koehorst et al., 2004). The tilt angle of the transmembrane helix $\theta = 18°$ is somewhat smaller than the value $\theta = 26°$ arising from solid state NMR (Marassi and Opella, 2003). However it is within the range of $20 \pm 10°$ as found earlier from solid state ^{13}C NMR (Glaubitz et al., 2000). From Stokes shift experiments a range of tilt angles from 18 to 28° was estimated (Koehorst et al., 2004). In this work (Koehorst et al., 2004) the tilt angle is given as a function of the distance between the AEDANS moiety and the α-helix axis. A tilt angle of $19 \pm 1°$ corresponding to the distance l_A = 9.5 Å, used in our work, is in excellent agreement with our value of 18±2°. The direction of the protein tilt ψ is the least sensitive parameter in our case. Nevertheless our value of $61 \pm 7°$ is close to 60° as found previously (Koehorst et al., 2004). This comparison shows that our model is performing well, certainly by taking into account that only three different mutants were used.

From Fig. 2.5 it can be noticed that some fits are not ideal. The reason for these deviations between simulated and experimental efficiencies could be related to the fact that the long AEDANS label arm is mobile within a restricted space angle, which size and direction differs for different mutants (Vos et al., 2005). A future enhancement of the model could be the implementation of the entire AEDANS conformational space for each mutant instead of assuming a constant acceptor arm normal to the helix axis. A further improvement of the precision of our model can be achieved by using the fluorescent data of the AEDANS (Koehorst et al., 2004) in a general global optimization algorithm. We are currently working on these challenging ideas.

The methodology developed here is not limited to M13 major coat protein and can be used in principle to study the bilayer embedment and structure of any α-helical single transmembrane protein (or peptide), and with some adaptations to transmembrane domains of larger membrane proteins. For example, the method was successfully applied to study the

aggregation of various WALP peptides in lipid bilayers of different thickness (Sparr et al., 2005).

Acknowledgments. This work was supported by contract no. QLG-CT-2000-01801 of the European Commission (MIVase – New Therapeutic Approaches to Osteoporosis: targeting the osteoclast V-ATPase). We would like to thank Ruud B. Spruijt for the preparation of the protein mutants and helpful comments on the work.

2.6. Appendix A. Derivation of energy transfer efficiency

Consider a protein-lipid system containing two types of fluorescently labeled proteins – with a single donor (denote it's quantity by C_u) and with a donor and acceptor (denote the quantity by C_l). Let us introduce two efficiencies of energy transfer E_u and E_l, characterizing energy transfer for the first and second protein population. The total energy transfer efficiency E for the system is then given by

$$E = \frac{C_u}{C_u + C_l} E_u + \frac{C_l}{C_u + C_l} E_l \tag{2.9}$$

Consider now the acceptor excitation spectrum for such a system in a general case

$$F^\lambda = \left(\Theta_A^\lambda + \Theta_u^\lambda + \Theta_l^\lambda\right)\gamma, \tag{2.10}$$

where Θ_A is the direct acceptor excitation at wavelength λ, Θ_u^λ the excitation due to energy transfer from unlabeled proteins, Θ_l^λ the excitation caused by energy transfer from labeled proteins (both intra and intermolecular), and γ is a constant that depends on the apparatus and experimental conditions. Taking into account the extinction coefficients of donor and acceptor and protein quantities this equation can be rewritten in the following form

$$F^\lambda = \gamma\left(C_l \varepsilon_A^\lambda + C_u \varepsilon_D^\lambda E_u + C_l \varepsilon_D^\lambda E_l\right), \tag{2.11}$$

where ε_A^λ is the extinction coefficient of acceptors at wavelength λ, and ε_D^λ the extinction coefficient of the donors. At λ = 290 nm the extinction coefficients are non-zero both for our donor (Trp-26) and acceptor (AEDANS). However at λ = 340 nm ε_D^{340} = 0. Taking into account the fluorescence at these two wavelengths and expressing the partial efficiencies via Eq. 2.9, the following descriptions for the fluorescence of the protein-lipid system can be obtained:

$$F^{290} = \gamma\left(C_l\varepsilon_A^{290} + \varepsilon_D^{290}\left(C_u + C_l\right)E\right), \tag{2.12}$$

$$F^{340} = \gamma C_l\varepsilon_A^{340}. \tag{2.13}$$

Dividing Eq. 2.12 by 2.13 and making simple rearrangements the following equation is obtained:

$$\left(\frac{F^{290}}{F^{340}} - \frac{\varepsilon_A^{290}}{\varepsilon_A^{340}}\right)\frac{\varepsilon_A^{340}}{\varepsilon_D^{290}} = \left(1 + \frac{C_u}{C_l}\right)E. \tag{2.14}$$

By introducing the ratio of the number of unlabelled to labeled proteins, r_{ul}, Eq. 2.14 can then be presented in the following form:

$$E = \frac{1}{1 + r_{ul}}\left(\frac{F^{290}}{F^{340}} - \frac{\varepsilon_A^{290}}{\varepsilon_A^{340}}\right)\frac{\varepsilon_A^{340}}{\varepsilon_D^{290}}. \tag{2.15}$$

2.7. Appendix B. Analytical equation for FRET in systems of M13 coat protein proteins incorporated into a lipid bilayer

Consider a system of labeled and unlabeled M13 coat protein incorporated into a lipid bilayer. Let r_{ul} be the molar ratio of the labeled and unlabeled proteins. The time decay of the fluorescence intensity, $\rho(t)$, of the donor in this system can then be described by:

$$\rho(t) = \frac{r_{ul}}{1 + r_{ul}}\rho_u(t) + \frac{1}{1 + r_{ul}}\rho_l(t), \tag{2.16}$$

where ρ_u is the fluorescence decay of the unlabeled proteins, and ρ_l the fluorescence decay of the labeled proteins. The coefficients in front of $\rho_{u,l}$ in Eq. 2.16 are the fractions of labeled and unlabeled proteins expressed in terms of r_{ul}.

The fluorescence decay of donors attached to unlabeled proteins ρ_u is affected by acceptors of other proteins, distributed around. For labeled proteins the intramolecular energy transfer should be taken into account as well. Thus:

$$\rho_u(t) = \rho_D(t) \times q_\sigma(t, R_0, \sigma, h, D_P) \tag{2.17}$$

$$\rho_l(t) = \rho_D(t) \times q_\sigma(t, R_0, \sigma, h, D_P) \times q_{intra}(t), \tag{2.18}$$

where ρ_D is the fluorescence of a single donor, q_σ the quenching effect by distributed acceptors, and q_{intra} the quenching effect by intramolecular energy transfer in labeled proteins. We assume now that the donor fluorescence has a single lifetime and can be described by

$$\rho_D(t) = \exp\left(-t/\tau_D\right), \tag{2.19}$$

where τ_D is a single donor lifetime. Alternatively, all expressions presented below may easily be reproduced for multiexponential donor fluorescence (Loura et al., 2001).

The quenching by intramolecular energy transfer is given by

$$q_{intra}(t) = \exp\left(-t\left(R_0/R_{intra}\right)^6/\tau_D\right), \tag{2.20}$$

where R_{intra} is the intramolecular donor-acceptor distance. Consider now the quenching due to distributed acceptors. The overall surface density of acceptors is given by

$$\sigma = \frac{N_A}{S} = \frac{N_l}{S_L N_L/2 + S_P(N_l + N_u)}, \tag{2.21}$$

where S is the area of the entire membrane, N_A the number of acceptors in the system, N_l the number of labeled proteins, N_u the number of unlabeled proteins, N_L the number of lipids, S_L the area occupied by a single lipid molecule, and S_P the area occupied by a single protein molecule. Taking into account the definitions of r_{LP} and r_{ul}, and considering cylindrical proteins, Eq. 2.21 can be presented in the form

$$\sigma = \left[\left(S_L r_{LP}/2 + S_P\right)\left(1 + r_{ul}\right)\right]^{-1} = 2\left[\left(S_L r_{LP} + D_P^2\pi/2\right)\left(1 + r_{ul}\right)\right]^{-1}. \tag{2.22}$$

Because of the possibility of parallel and anti-parallel protein orientations, the initial acceptor density σ is divided over the two leaflets. For each leaflet the acceptor density σ_1 is given by:

$$\sigma_1 = \sigma/2 = \left[\left(S_L r_{LP} + D_P^2\pi/2\right)\left(1 + r_{ul}\right)\right]^{-1}. \tag{2.23}$$

Donors are divided over the two leaflets as well. The symmetry of the system then leads to an equivalence of relative distances between each donor plane and two acceptor planes. Therefore the system can be substituted with a system containing one layer of donors and two layers of acceptors at the distances:

$$h_{\mathrm{I}} = \left|Z_D - Z_A\right|, \text{ and } h_{\mathrm{II}} = \left|Z_D + Z_A\right| \tag{2.24}$$

where Z_D and Z_A are the z coordinates (in the membrane axis system) of a donor and acceptor, respectively, attached to a protein with an "upright" orientation.

The analytical solution for the donor fluorescence decay in the presence of uniformly distributed acceptors in a plane was given by Davenport (Davenport et al., 1985). Taking into account two layers of acceptors located at h_{I} and h_{II}, the quenching effect on the donor fluorescence is given as follows:

$$\begin{aligned} q_\sigma(t) &= q_\sigma{}^{\mathrm{I}}(t) \times q_\sigma{}^{\mathrm{II}}(t) = \\ &= \exp\left\{ -2\pi\sigma_1 h_{\mathrm{I}}{}^2 \int_{\alpha=0}^{\frac{h_{\mathrm{I}}}{\sqrt{h_{\mathrm{I}}{}^2 + D_P^2}}} \left(1 - \exp\left[-\frac{t}{\tau_D} \left(\frac{R_0 \alpha}{h_{\mathrm{I}}} \right)^6 \right] \right) \alpha^{-3} d\alpha \right\} \times \\ &\times \exp\left\{ -2\pi\sigma_1 h_{\mathrm{II}}{}^2 \int_{\beta=0}^{\frac{h_{\mathrm{II}}}{\sqrt{h_{\mathrm{II}}{}^2 + D_P^2}}} \left(1 - \exp\left[-\frac{t}{\tau_D} \left(\frac{R_0 \beta}{h_{\mathrm{II}}} \right)^6 \right] \right) \beta^{-3} d\beta \right\} \end{aligned} \tag{2.25}$$

The energy transfer efficiency can be calculated using the relative integrated fluorescence intensity of the donors in the presence and absence of acceptors:

$$E = 1 - \frac{\int_0^\infty \rho(t) dt}{\int_0^\infty \rho_D(t) dt}. \tag{2.26}$$

The integrated fluorescence of a single donor in the case of one exponential decay equals to τ_D. After substitution of ρ, the energy transfer efficiency E can be expressed in terms of ρ_D, q_σ, and q_{intra} as follows:

$$E = 1 - \frac{1}{\tau_D} \int_0^\infty \rho_D(t) \times q_\sigma(t) \times \frac{r_{ul} + q_{intra}(t)}{1 + r_{ul}} dt. \tag{2.27}$$

2.8. Appendix C. Analysis of the solutions obtained by SBF

The FRET model that is used in our SBF fitting has a random nature and therefore the error function χ^2 (Eq. 2.8) is a stochastic one. To deal with this stochastic effect, the fitting procedure needs to be performed several times (we take 100, which is found to be sufficiently large) with different starting fitting parameters. This approach results in a distribution of solutions and each of the resulting solutions has a different χ^2 value. A typical distribution of resulting χ^2 values is shown in Fig. 2.7 A.

In this case, the selection of the parameter set corresponding to the minimal χ^2 is not statistically correct, because a low χ^2 can be the result of a random deviation. At the same time, averaging of all solutions found will lead to an incorrect result as well, because many solutions with a high χ^2 are included. These solutions do not show a reasonable fit between the modeled and experimental data and appear only because the optimization algorithm is falling into false local minima.

To reduce the randomness of a single solution and to find the best solution in the parameter space, we use the following approach, which is often found in evolutionary computing (Strancar et al., 2005). A part of the solutions with the lowest χ^2 values is selected. This corresponds to selecting the quantile χ^2_q of the χ^2 distribution. The solutions with χ^2 less than the selected quantile are considered as an "elite" subset, and the mean value of the parameters inside this "elite" subset is then taken as the result of the fitting. The problem of this approach now reduces to finding the optimal size for the "elite" solutions, i.e. the value q in the quantile χ^2_q.

The selection of the optimal q is a problem-related task and cannot be analytically solved in general. Therefore we employ an empirical approach. Using our numerical model the analogues of experimental data were simulated for a known parameter vector $\boldsymbol{P}$, and these synthetic data were fitted by the same model. The resulting solutions were analyzed using the quantile approach with various values of q. This provides the resulting parameter vector $\boldsymbol{P}^*$. To validate the precision of the representative solutions found, we introduce a function e, which is the sum of the parameter deviations:

$$e = \sum_{i=1}^{n_p} \left(\frac{P_i^* - P_i}{P_i} \right)^2 , \tag{2.28}$$

where n_p is the number of parameters, and P_i is i-th parameter from the parameter vector $\boldsymbol{P}$. The sum parameter deviation e is related to the inaccuracy in the resulting parameters. The behavior of the function e with respect to q for our FRET model is depicted in Fig. 2.7 B. On increasing q the error is decreasing, as would be expected, since the "noise" is reduced. However after taking more solutions into account, the error is increasing again, because bad solutions are coming in. The minimal deviation in the parameters is reached for q = 20%.

To be fully applicable, the algorithm needs all "elite" solutions belonging to the neighborhood of a single χ^2 minimum, and their differences should be caused by simulation randomness. If the solutions would form several separated clusters, the same approach should be applied to each of those clusters, and the solutions found should be considered as possible states for the system. However, this does not happen in our case. The additional advantage of the proposed algorithm is that it gives direct insight in the error range. The standard deviation of parameters inside the "elite" subset of solutions therefore can be used as a characteristic of the error range of the resulting solution.

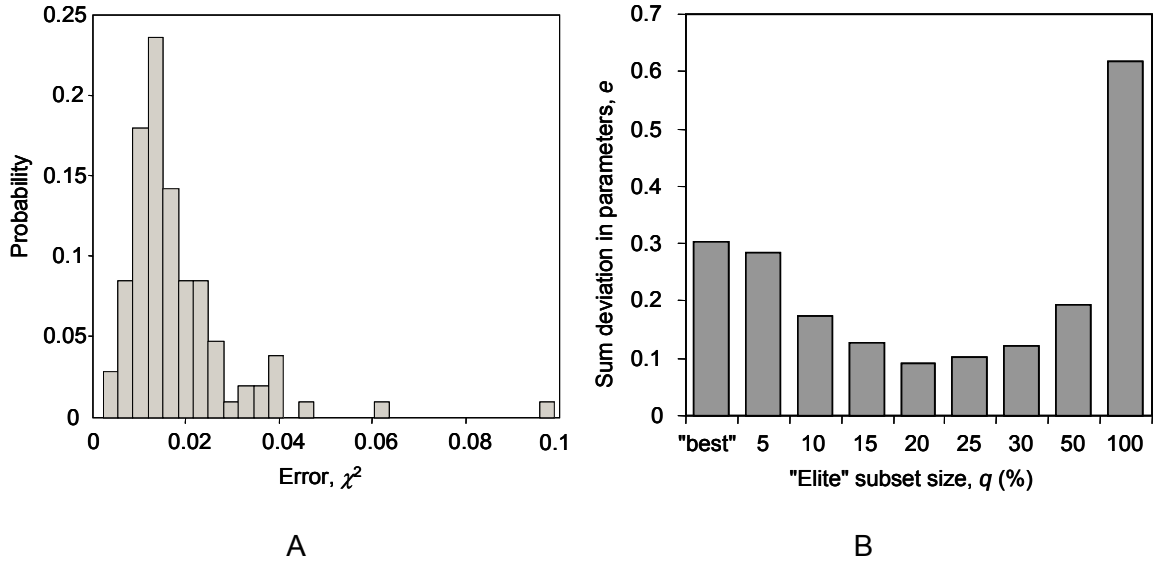


Figure 2.7. (A) Distribution of χ^2 of the solutions found after 100 runs of SBF on experimental data and (B) behavior of the sum parameter deviation e (Eq. 2.28) with respect to the "elite" subset size q. The result in (B) is obtained after averaging the results of three independent numerical simulations. For all of them the optimal q was around 20%.

3. FRET STUDY OF MEMBRANE PROTEINS: DETERMINATION OF THE TILT AND ORIENTATION OF THE N-TERMINAL DOMAIN OF M13 MAJOR COAT PROTEIN

Petr V. Nazarov, Rob B.M. Koehorst, Werner L. Vos, Vladimir V. Apanasovich, Marcus A. Hemminga

Accepted for publication by *Biophysical Journal*, **2007**, 92(4).

ABSTRACT

Formalism for membrane protein structure determination was developed. This method is based on steady-state Förster resonance energy transfer (FRET) data and information about the position of the fluorescence maxima on site-directed fluorescent labeled proteins in combination with global data analysis utilizing simulation-based fitting. The methodology was applied to determine the structural properties of the N-terminal domain of the major coat protein from bacteriophage M13 reconstituted into unilamellar DOPC:DOPG (4:1 mol/mol) vesicles. For our purpose, the cysteine mutants A7C, A9C, N12C, S13C, Q15C, A16C, S17C and A18C in the N-terminal domain of this protein were produced and specifically labeled with the fluorescence probe AEDANS. The energy transfer data from the natural Trp26 to AEDANS were analyzed assuming a two-helix protein model. Furthermore, the polarity Stokes shift of the AEDANS fluorescence maxima is taken into account. As a result a low-resolution structure of the protein was obtained, showing an overall α-helical protein conformation, close to the protein conformation in the intact phage, which is tilted by 18° with respect to the normal to the membrane. The methodology developed here is not limited to the M13 major coat protein and can be used in principle to study the bilayer embedment and structure of any protein for which a one or two-helix model can be applied.

3.1. Introduction

Understanding of disease pathways and developments for novel drugs are impossible without fundamental knowledge about structure and functionality of membrane proteins (Hesselink et al., 2005; Karmazyn et al., 2005; Vassar, 2002; Weinglass et al., 2004).

Membrane proteins represent more than 50% of all present and future drug targets. High-resolution structural studies of membrane proteins by X-ray crystallography or by nuclear magnetic resonance (NMR) spectroscopy have traditionally been limited by technical and practical difficulties (Torres et al., 2003). This makes structure determination of membrane proteins still a key challenge in structural biology. Therefore new biophysical characterization techniques are needed to advance the field. Recently, we have shown that steady-state Förster (or fluorescence) resonance energy transfer (FRET) provides an attractive alternative (Nazarov et al., 2004; Nazarov et al., 2006; Vos et al., 2005). In this work, we tested the FRET methodology on bacteriophage M13 major coat protein incorporated in membranes. Apart from obtaining structural information about the lipid-bound state of the coat protein, by using FRET also significant progress was obtained in the understanding the membrane embedment of the protein (Nazarov et al., 2006), the protein-protein and protein-lipid interactions (Fernandes et al., 2004; Fernandes et al., 2003).

M13 major coat protein, as well as coat proteins from related filamentous bacteriophages, has been used extensively as a biophysical reference system for studying membrane protein embedment (for a review see: (Stopar et al., 2006a)). The coat protein is a small protein with a molecular weight of about 5240 Da, which forms a cylindrical shell around the DNA in the phage particle. In the phage particle the protein is largely α-helical with 4-5 flexible unstructured amino acid residues in the N-terminus (Marvin et al., 1994). Previously, it was assumed that the insertion of the protein in a lipid bilayer is accompanied with a major structural rearrangement that splits the continuous α-helix in the phage particle into an N-terminal amphipathic and a transmembrane helix perpendicular to each other (Almeida and Opella, 1997; Bashtovyy et al., 2001; Bogusky et al., 1988; Bogusky et al., 1987; Henry and Sykes, 1992; Henry et al., 1987; Leo et al., 1987; Marvin, 1998; McDonnell et al., 1993; Wolkers et al., 1997). However, recent studies propose that the change in the secondary structure on insertion is not so dramatic and that the protein is still dominated by a high content of helical structure (Koehorst et al., 2004; Meijer et al., 2001b; Spruijt et al., 2000; Spruijt et al., 2004; Vos et al., 2005). Therefore, the difference between the two models is mostly confined to the topology and orientation of the two helices.

Previously FRET was applied to AEDANS-labeled cysteine mutants of M13 major coat protein reconstituted into vesicles. The energy transfer data from the natural tryptophan at position 26, which is used as a donor, to AEDANS were analyzed assuming a helix model for the transmembrane domain of the protein (Nazarov et al., 2006). The method allowed the determination of the depth, tilt angle, and direction of tilt of the protein in the membrane. To

resolve the problem concerning the relative orientation of the N-terminal and transmembrane domain, we present here an extension of this approach, by introducing a two-helix model describing the N-terminal and transmembrane helix domains of the M13 coat protein. This methodology results in a low-resolution structure of the entire protein, including the tilt and orientation of the N-terminal domain with respect to the transmembrane domain.

3.2. Experimental

3.2.1. Sample preparation

As in the previous study (Nazarov et al., 2006), the lipid bilayer systems were prepared from dioleoylphosphatidylcholine (DOPC, 18:1PC) and dioleoylphosphatidylglycerol (DOPG) lipids in a 4:1 molar ratio, denoted as DOPC:DOPG. DOPC was purchased from Avanti Polar Lipids and DOPG was purchased from Sigma.

Site-specific cysteine mutants of M13 major coat protein were prepared, purified and labeled with AEDANS (Molecular Probes) as described previously (Spruijt et al., 2000). Wild-type protein and AEDANS-labeled M13 coat protein mutants were reconstituted into phospholipid bilayers as reported earlier (Spruijt et al., 1989).

Protein titration experiments were carried out using the same protocol as published recently (Nazarov et al., 2006). We used AEDANS-labeled cysteine mutants of M13 coat protein with the cysteine residue at positions 7 (A7C), 9 (A9C), 12 (N12C), 13 (S13C), 15 (Q15C), 16 (A16C), 17 (S17C) and 18 (A18C). Titration experiments were performed in which the wild-type protein concentration was increased, whereas the mutant concentration was kept constant. The sample conditions for these titrations are given in Table 3.1. The labeling efficiencies were determined as reported previously (Spruijt et al., 1996) and are given in Table 3.1. The labeling efficiency is explicitly taken into account in Table 3.1 in the ratio of the number of unlabeled to labeled proteins (r_{ul}), as it affects the acceptor concentration and therefore the energy transfer efficiency.

For the fluorescence experiments stock solutions of protein mutants and wild-type protein solubilized in cholate buffer were mixed with solutions of lipids in the same buffer, as described previously (Spruijt et al., 1989). Repeated dialysis of the mixtures in cholate-free buffer was performed to remove the cholate in the sample. The lipid loss during dialysis can vary between 20-30% (Spruijt et al., 1989, Nazarov, 2006 #56), and this fact is accounted for in the analysis of the experimental data.

3.2.2. Fluorescence experiments

Optical spectroscopy. Fluorescence emission and fluorescence excitation measurements were performed on a Fluorolog 3.22 manufactured by Jobin Yvon-Spex as described elsewhere (Gustiananda et al., 2004; Nazarov et al., 2006; Vos et al., 2005). The position of the AEDANS emission maximum was different for different labeled mutants, because the Stokes shift of AEDANS fluorescence significantly depends on the local polarity of the environment of the label, and thus on the distance between the label and the center of the lipid bilayer (Koehorst et al., 2004; Spruijt et al., 2004). During emission detection AEDANS was excited at 365 nm, and the fluorescence was measured between 400 and 600 nm. Emission spectra were corrected for background fluorescence using equimolar solutions of lipid vesicles with incorporated wild type (no AEDANS present) proteins. The positions of the AEDANS fluorescence maxima for the various mutants are given in Table 3.1.

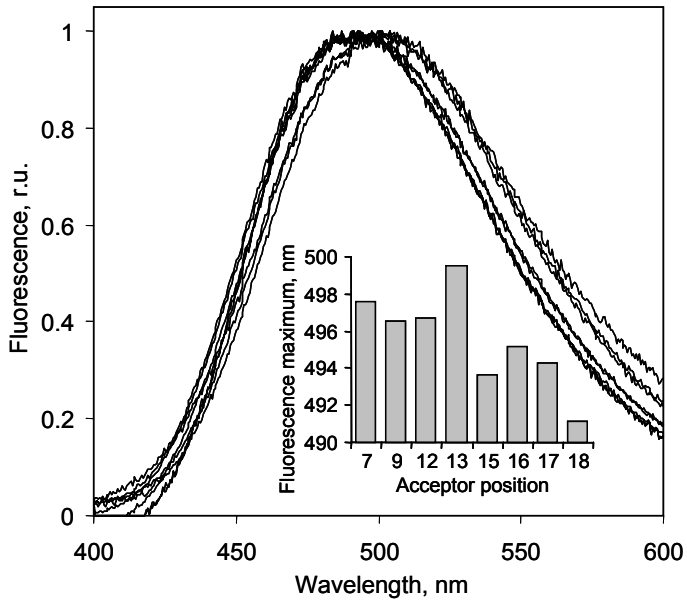
For fluorescence excitation measurements the detection wavelength was set at the maximum of the acceptor (AEDANS) fluorescence of a particular mutant and the excitation wavelength was scanned from 260 to 400 nm. The resulting emission AEDANS spectra for all mutants, and examples of excitation spectra for mutant N12C are presented in Fig. 3.1 A and B, respectively.

The applied slit widths of the detection and excitation monochromators corresponded to 5 and 2 nm band pass, respectively. The spectra were automatically corrected on the Fluorolog 3.22 for variations in the lamp output by dividing the sample signal by that of an internal reference detection system. All excitation spectra were corrected for background fluorescence using equimolar solutions of lipid vesicles with incorporated wild type proteins. The detected fluorescence exclusively belongs to AEDANS (Nazarov et al., 2006). The temperature during all measurements was 20°C. Because of the small protein concentrations used in our experiments (about 1 μM), errors caused by inner filter effects can be neglected.

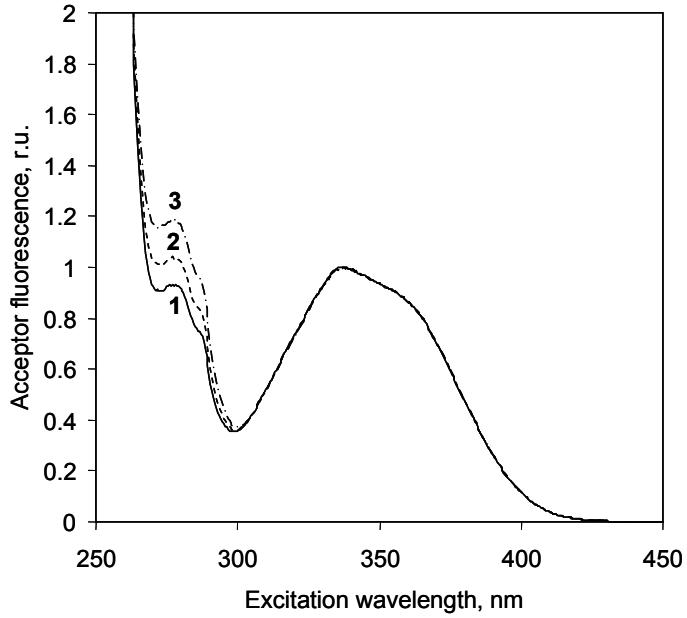
Analysis of AEDANS emission and excitation spectra. The position of the AEDANS emission maxima was determined using a polynomial approximation of the top part of the emission peak as in (Koehorst et al., 2004).

Table 3.1. Sample composition of M13 major coat protein incorporated into DOPC:DOPG vesicles given in terms of r_{LP} and r_{ul}, labeling efficiencies, and observed acceptor fluorescence maxima and energy transfer efficiencies E for mutants with acceptor positions n_A at 7, 9, 12, 13, 15, 16, 17, and 18.

Data set #	1	2	3	4	5	6	7	8
Mutant	A7C	A9C	N12C	S13C	Q15C	A16C	S17C	A18C
n_A	7	9	12	13	15	16	17	18
Acceptor fluorescence max. λ_{max}	497.6 nm	496.5 nm	496.7 nm	499.5 nm	493.6 nm	495.1 nm	494.2 nm	491.1 nm
Labeling efficiency	0.44	0.78	0.79	0.55	0.53	0.85	0.54	0.56
r_{LP}	336.0	217.0	276.9	561.5	422.6	239.7	267.7	248.6
r_{ul}	1.27	0.28	0.27	0.82	0.89	0.18	0.85	0.79
E	0.172	0.463	0.505	0.338	0.443	0.880	0.448	0.488
r_{LP}	213.2	158.2	184.8	286.2	245.1	169.9	180.7	171.8
r_{ul}	2.58	0.76	0.90	2.57	2.25	0.66	1.74	1.58
E	0.119	0.366	0.360	0.184	0.261	0.650	0.323	0.354
r_{LP}	156.2	124.5	138.7	192.0	172.6	131.6	136.3	131.2
r_{ul}	3.89	1.24	1.53	4.32	3.62	1.14	2.64	2.38
E	0.095	0.307	0.279	0.129	0.200	0.513	0.261	0.286
r_{LP}	123.2	102.6	111.0	144.5	133.2	107.4	109.5	106.2
r_{ul}	5.20	1.71	2.16	6.07	4.99	1.63	3.53	3.18
E	0.084	0.260	0.238	0.102	0.152	0.436	0.217	0.248
r_{LP}	101.7	87.2	92.5	115.8	108.5	90.7	91.5	89.1
r_{ul}	6.51	2.19	2.79	7.82	6.35	2.11	4.42	3.98
E	0.076	0.231	0.199	0.086	0.129	0.383	0.191	0.225
r_{LP}	86.6	75.9	79.3	96.6	91.5	78.5	78.5	76.8
r_{ul}	7.81	2.67	3.42	9.57	7.72	2.59	5.31	4.78
E	0.067	0.207	0.184	0.075	0.111	0.340	0.175	0.198
r_{LP}	57.0	52.2	52.9	61.2	59.1	53.4	52.6	51.8
r_{ul}	12.4	4.33	5.63	15.69	12.50	4.28	8.43	7.57
E	0.055	0.160	0.137	0.055	0.079	0.254	0.138	0.149



A



B

Figure 3.1. (A) Emission spectra of M13 protein mutants A7C, A9C, N12C, S13C, Q15C, A16C, S17C, and A18C with AEDANS-labeled Cys after subtraction of the fluorescence of equimolar wild-type samples. The histogram shows the values of the acceptor emission maxima of the mutants. (B) Experimental excitation spectra (detected at 496 nm) obtained for mutant N12C at three titration points of wild-type proteins. Labels 1 to 3 correspond to r_{ul} values of 0.27, 2.16, and 5.63, respectively. The lipid-to-protein ratios r_{LP} are 277, 111, and 53 (see Table 3.1). The sample showing the highest peak at 290 nm (spectrum 3) has the highest protein density (lowest r_{LP}) and r_{ul}. Although the efficiency of energy transfer (Fig. 3.3) for this case is smallest, the overall energy absorbed by the donors in such a system, and therefore the transferred (intermolecular), is higher than for the other values of r_{LP} and r_{ul} (Nazarov et al., 2006).

The derivation of the mathematical expressions for the analysis of the experimental excitation spectra is given in the previous work (Nazarov et al., 2006). For our analysis we used the energy transfer efficiency E, which can be calculated from the fluorescence intensities (Lakey et al., 1993; Nazarov et al., 2006) by

$$E = \frac{1}{1+r_{ul}}\left(\frac{F^{290}}{F^{340}} - \frac{\varepsilon_A^{290}}{\varepsilon_A^{340}}\right)\frac{\varepsilon_A^{340}}{\varepsilon_D^{290}}, \tag{3.1}$$

where r_{ul} is the ratio of the number of unlabeled to labeled proteins. For every sample the ratio of the fluorescence intensity at 290 nm, F^{290}, (mainly donor excitation) to that at 340 nm, F^{340}, (exclusively acceptor excitation) was calculated, being a measure of the donor-to-acceptor energy transfer. The ratio F^{290}/F^{340} was corrected for direct excitation of AEDANS at 290 nm by subtracting the ratio of the extinction coefficients $\varepsilon_A{}^{290}/\varepsilon_A{}^{340}$ = 0.20 (this ratio was calculated using mutant Y21A/Y24A/W26A/G23C). Finally, the ratio of the extinction coefficients of the acceptor at 340 nm ($\varepsilon_A{}^{340}$) and donor at 290 nm ($\varepsilon_D{}^{290}$) have to be taken into account in Eq. 3.1 ($\varepsilon_A{}^{340}/\varepsilon_D{}^{290}$ = 1.2).

3.2.3. Förster distance

The value of Förster distance R_0, needed for simulation of energy transfer, was calculated using Eq. 3.2

$$R_0 = 9780\left(\kappa^2 n^{-4} Q_D J\right)^{1/6}. \tag{3.2}$$

In this equation Q_D = 0.23 was taken, which is the quantum yield of tryptophan in dimyristoyl phosphatidylcholine (DMPC) bilayers (Fisher and Ryan, 1999). The overlap integral J is calculated from the emission spectrum of the wild-type protein and the absorption spectrum of the AEDANS-labeled Y21A/Y24A/W26A/G23C mutants, which had no tryptophan at position 26. This results in a value of 5.96×10^{-15} $M^{-1}cm^3$. The orientation factor κ^2 is approximated by its isotropic dynamic average, giving a value of 2/3 (Kamal and Behere, 2002; Lakshmikanth et al., 2001; Loura et al., 1996; Nazarov et al., 2006; Vos et al., 2005). For simplicity the refractive index of the medium is assumed to be constant, and equal to 1.4 (Davenport et al., 1985; Lakowicz, 1999). These parameters result in a Förster radius R_0 of 24 Å. It should be noted, that the excitation band of AEDANS with its maximum around 340 nm does not change with the position of the labeled cysteine. This implies that the Förster distance for the donor-acceptor pair is equal for all mutants.

3.3. Methodology

3.3.1. Model for M13 major coat protein incorporated into a lipid bilayer

In this paper, we will extents our previous single-helix model for the M13 major coat protein (Nazarov et al., 2006) to a two-helical model. This model consists of two flexibly linked helical domains connected via a kink (Fig. 3.2): one domain reflects the transmembrane protein part and the other domain the N-terminal protein part that is supposed to stick out of the membrane (Glaubitz et al., 2000; Koehorst et al., 2004; Marassi and Opella, 2003; Stopar et al., 2006a; Stopar et al., 2003). The conformation of each domain is assumed to be a perfect α-helix. The main axis of the protein *O* is parallel to the transmembrane protein domain and defines the *z* axis of the axes system of the protein. The orientation of the *x* axis is defined by the location of the donor Trp26, which is used as the reference amino acid residue. The complete set of structural parameters that determines the location and conformation of the protein is presented in Table 3.2. Some of the protein parameters related to position of the transmembrane domain are described in the previous study (Nazarov et al., 2006).

The parameter ranges given in Table 3.2 indicate the range of values considered in the simulations. It should be noted that the two-model could be easily generalized to other membrane proteins. Furthermore, a one-helix protein is a special case of the two-helix model with $\varphi = \omega = 0^{\circ}$.

The two-helix protein model is incorporated in a membrane as described previously (Nazarov et al., 2004; Nazarov et al., 2006). A square region of a bilayer containing a certain number of randomly incorporated proteins (N_P) is considered. By using a three-dimensional mathematical description, protein molecules as shown in Fig. 3.2 are inserted randomly (both in location as well as in orientation) into the lipid bilayer in the way that the angle θ between the membrane normal and their main axis *O* of the transmembrane domain is between 0 and 90º. The direction of the protein tilt is given by ψ. A value $\psi = 0$ means that protein is tilted towards the C_α of the reference (n_0) amino acid residue. The depth of protein insertion is given by parameter *d*. It is assumed, that when inserted into the membrane, the proteins occupy a cylindrical region in both bilayer leaflets with a protein exclusion distance D_P. Within this region, no lipids or other proteins can be located.

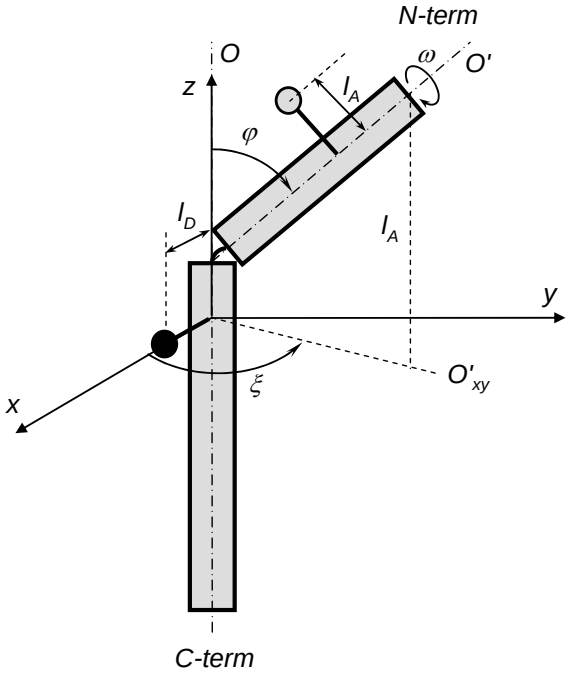


Figure 3.2. Schematic drawing of the two-helix protein model with a donor (Trp26, black circle, located at a distance l_D from the protein helix axis) and acceptor (AEDANS, gray circle, located at a distance l_A from the protein helix axis) attached at positions 26 and 9, respectively, in its own protein axis system (x, y, z). The orientation of the x axis is defined by the location of Trp26, which is used as the reference amino acid residue. The complete set of structural parameters that describes the protein-lipid system is presented in Table 3.2.

The area of the considered square region of the membrane is calculated from the experimental L/P ratio (r_{LP}), the protein exclusion distance (D_P), the area per two lipid molecules (S_L) and the ratio of lipids lost during dialysis to their initial quantity (i.e. the lipid loss L) in the following way:

$$S = N_P\left(S_L r_{LP}(1-L)/2 + \pi D_P^2/4\right). \tag{3.3}$$

Furthermore, to be able to work with mixtures of labeled and unlabelled protein molecules, the ratio r_{ul} between the number of unlabeled and labeled proteins is introduced into the model.

Similar to the experimental reconstituted protein-lipid system, protein molecules can be inserted into the model membrane randomly with “parallel” and “anti-parallel” orientations; this means that the N-terminal domain of the protein can be located either in the upper or in the lower leaflet of the membrane with equal probabilities. The result of these equiprobable orientations is that the membrane system contains two layers of donors and two layers of acceptors.

Table 3.2. Definition of the parameters used in the two-helix model of proteins embedded in lipid bilayers. In the simulations the parameters n_k, φ, ω, ξ, and L are varied. Parameters n_A, r_{LP} and r_{ul} are determined by the experiment; the other parameters are taken from previous work (Nazarov et al., 2006) and are fixed as shown in the table.

Parameter	Range/Value	Unit	Description
n_0	26	–	The position of a reference amino acid residue. The projection of its C_α to the helix axis of the protein *O* gives the origin of the coordinate system of the protein. Position n_0 = 26 was selected for the transmembrane domain of M13 major coat protein.
h	1.5	Å	Translation per amino acid residue along the helix; this is 1.5 Å for a perfect α-helix.
n_r	3.6	–	Number of amino acid residues per one turn; this is 3.6 for a perfect α-helix.
n_D	26	–	Donor position; position of amino acid residue given by the donor. For M13 coat protein the donor is Trp-26, which is located in the transmembrane domain.
n_A	1 – 50	–	Acceptor position; position of amino acid residue labeled by the acceptor. For the transmembrane domain of M13 coat protein the acceptor positions are 24, 38 and 46.
l_D	6.5	Å	Donor arm, the average distance from the donor moiety to the helix axis. A value l_D = 6.5 Å was taken (Koehorst et al., 2004).
l_A	9.5	Å	Acceptor arm, the average distance from the acceptor moiety to the helix axis. A value l_A = 9.5 Å was taken (Koehorst et al., 2004).
n_k	1 – 25	–	Position of helix kink; position of amino acid residue from which the N-terminal helix starts.
θ	18	°	Protein tilt angle; the angle between the helix axis and the normal to the membrane. The value of 18° is found in the previous study (Nazarov et al., 2006).
d	8.5	Å	Distance from the origin of the coordinate system of the protein to the centre of the bilayer is 8.5 Å (Nazarov et al., 2006).
ψ	60	°	Protein tilt direction; the direction of the protein transmembrane domain tilting is ~60° as found earlier (Koehorst et al., 2004; Nazarov et al., 2006).
N_P	500	–	Number of proteins in the system. All simulations were performed for models containing 500 proteins.
S_L	72	$Å^2$	Area occupied by a lipid in one leaflet of a bilayer; the average area for the DOPC:DOPG system is 72 $Å^2$ (Fernandes et al., 2003).
L	0.0 – 1.0	–	Lipid loss; ratio of lipids lost during dialysis to their initial quantity.

Parameter	Range/Value	Unit	Description
D_P	10	Å	Protein exclusion distance; minimal protein-protein distance. For M13 coat protein a value D_P = 10 Å was taken.
r_{LP}	≥ 0	–	Lipid to protein ratio.
r_{ul}	≥ 0	–	Ratio between the number of unlabeled and labeled proteins.
k	0	–	Protein-protein association probability, defined as the percentage of clustered proteins with respect to the total number of proteins, for considered case ~0 (Fernandes et al., 2004 ; Nazarov et al., 2006) .
R_0	24	Å	Förster distance. A value of 24 Å is calculated using the data about the photophysical properties of the donor and acceptor.
φ	0 – 90	°	N-terminal helix tilt angle; the angle between protein main axis and the N-terminal helix main axis.
ξ	–180 – 180	°	N-terminal helix tilt direction; the direction of the N-terminal helix with respect to the x axis of the protein axis system.
ω	–180 – 180	°	N-terminal helix coaxial rotation; the turning angle of N-terminal helix around its main axis defining the direction of amino acid residues (towards water or lipid phase). The case $\omega = 0°$ corresponds to an ideal α-helix, bent at position n_k by angle φ.

A protein-protein association probability k can in principle be introduced to take into account the ability of the membrane proteins to form oligomers or clusters (Nazarov et al., 2006). However for the present study this value is about 0 and can be neglected. Therefore the distribution of proteins in the bilayer is considered as uniformly random.

Apart from the structural parameters and parameters related to the composition of the protein-lipid system, the Förster distance R_0 is introduced in the calculations for the energy transfer, as described below.

3.3.2. Models for FRET

Being in an excited state a fluorescent molecule has a dipole-dipole interaction with other molecules in close proximity, which can lead to energy transfer from the excited molecule to the non-excited ones. If we assume that the emission spectrum of the donor overlaps with the absorption spectrum of the acceptor, the photon absorbed by the donor can be transferred to the acceptor with a rate constant k_{ET} depending on the sixth power of the distance between the donor and acceptor

$$k_{ET} = \frac{1}{\tau_D}\left(\frac{R_0}{R}\right)^6, \tag{3.4}$$

where τ_D is the lifetime of an isolated donor, R the distance between the donor and acceptor. The Förster distance R_0 is given by Eq. 3.2.

Consider now a system of multiple donors and acceptors that are fixed at their positions. Let us number the donors i=1..N_D, and acceptors j=1..N_A. Here N_D is the number of donor molecules, and N_A the number of acceptor molecules. The probability for each donor to transfer energy to one of the acceptors can then be calculated as follows:

$$p_i = \frac{\sum_{j=1}^{N_A} k_{i,j}}{\frac{1}{\tau_D} + \sum_{j=1}^{N_A} k_{i,j}} = \frac{\sum_{j=1}^{N_A} \left(R_0 / R_{i,j}\right)^6}{1 + \sum_{j=1}^{N_A} \left(R_0 / R_{i,j}\right)^6}, \tag{3.5}$$

where $R_{i,j}$ is the distance between the i-th donor and j-th acceptor.

The mean probability of energy transfer events for all donor molecules gives the energy transfer efficiency E for the entire system:

$$E = < p_i >_{N_D}. \tag{3.6}$$

To analyze the experimental steady-state fluorescence data for our system, a steady-state FRET simulation is employed as described by (Nazarov et al., 2006). The simulation starts with the generation of the spatial model for the protein-lipid system. This model provides the coordinates of each donor and acceptor. The energy transfer efficiency E is then calculated using Eqs. 3.5 and 3.6. Because of the stochastic nature of the spatial model, the resulting energy transfer efficiency contains stochastic deviations. Therefore the simulations are executed several times (in our case: 50) to make the results statistically relevant.

3.3.3. Simulation-based fitting approach to experimental data analysis

As a measure of the goodness of the fit the following criterion was introduced:

$$\chi^2 = \sum_{i=1}^{N} \left(E_i^e - E_i^s\right)^2, \tag{3.7}$$

where N is the number of data points, E_i^e the experimentally obtained energy transfer efficiency, and E_i^s the simulated energy transfer efficiency. To fit the modeled energy transfer efficiencies to the experimental ones the Nelder-Mead “simplex” method (Nelder and Mead, 1965) is used. To increase the robustness of the method and the precision of the

solution a global analysis approach is chosen, and therefore all experimental data were fitted simultaneously (Beechem and Brand, 1986).

Because of the stochastic behavior of the FRET model, the error function χ^2 is stochastic as well, and the parameters obtained after each fit contain random deviations that are dependent on the sensitivity of the energy transfer to variations of the parameters. Therefore, to deal with this stochastic effect and to avoid possible local minima, the fitting procedure is performed a number of times with different initial estimations of the fitting parameters. The methodology used for the analysis of the resulting solutions and the selection of the representative solutions are recently described and discussed in (Nazarov et al., 2006).

All models were realized as C++ classes. The Borland C++ Builder 6.0 environment was used to combine the developed models, OpenGL visualization and simulation-based fitting algorithms into a software tool called FRETsim. The C++ classes and software are available from the authors upon request.

3.3.4. Handling of the Stokes shift information

The fluorescence emission of molecules in different solvents is significantly affected by the solvent polarity. The shift in fluorescence emission with respect to the absorption (and therefore the dependence of emission spectra with respect to polarity of the local environment) is called the Stokes shift (Lakowicz, 1999; Ren et al., 1999; Valeur, 2001). The dependency of the fluorescence emission maximum with respect to the polarity of the local environment of AEDANS-labeled cysteine mutants of M13 major coat protein incorporated in lipid bilayers was discussed recently (Koehorst et al., 2004; Spruijt et al., 2004). Therefore, we decided to use the Stokes shift information as an additional filtering for the structures obtained after fitting of the FRET data.

Unfortunately, analytical expressions, describing the behavior of the Stokes shift exist only for the internal hydrophobic part of lipid bilayers (Koehorst et al., 2004). However, a monotonic behavior of the polarity with respect to the absolute value of the *z*-coordinate of AEDANS in a bilayer system is demonstrated (White and Wimley, 1999). This result is probably related to the presence of motional averaging in the liquid crystalline phase (we are working with bilayer systems above the gel-to-liquid crystalline phase transition temperature). The possible effects of different polarity of neighbor amino acid residues can be neglected in our case, because of a long link between AEDANS moiety center and the protein backbone. This monotonic behavior enables us to build qualitative rules characterizing the relative *z*-coordinates for a polarity probe that can be applied for sites on the protein in the headgroup

region of the membrane or in the water phase. For example, consider two mutants with AEDANS emission maxima at wavelengths λ_1 and λ_1, and $\lambda_1<\lambda_2$. Consequently the relation for the z-coordinates of the fluorescent labels $|z_1| < |z_2|$, is also true. This relation can be considered as a qualitative rule: "the AEDANS position in the first mutant is closer to the membrane center then of the second mutant".

Three types of qualitative relations were selected to describe the positions of AEDANS in various mutants, each associated with a characterizing number $\in[-1, 0, 1]$. These numbers can be combined into a matrix $\boldsymbol{M}$, presenting the polarity rules for all mutants that are taken into account. The matrix elements M_{ij} describe the relation between the depth of i-th and j-th mutant. Assuming constant data precision for all mutants and denoting the maximal spread in the determined λ values as $\Delta\lambda$, the value of element M_{ij} is set according to the following scheme:

- if $\Delta\lambda_i - \Delta\lambda_j > \Delta\lambda \Rightarrow |z_i| > |z_j|$, $M_{ij} = 1$;
- if $|\Delta\lambda_i - \Delta\lambda_j| \le \Delta\lambda \Rightarrow |z_i| \approx |z_j|$, $M_{ij} = 0$;
- if $\Delta\lambda_i - \Delta\lambda_j < -\Delta\lambda \Rightarrow |z_i| < |z_j|$, $M_{ij} = -1$.

The resulting matrix is symmetric with zero diagonal elements.

To quantify the deviation between experimental relations and modeled ones, the following parameter is introduced

$$\delta = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left| M_{ij} - M_{ij}^{*} \right|, \tag{3.8}$$

where M_{ij}^{*} – is the matrix element describing relations obtained from the model of the protein for i-th and j-th mutant.

In our approach, the value of δ is used for an additional validation of the results coming out from the simulation-based fittings.

3.4. Results

3.4.1. Analysis of FRET data

We started the study of the protein structure with a simultaneous analysis of all eight data series, measured for AEDANS at positions 7, 9, 12, 13, 15, 16, 17, and 18. The best-achieved fit, characterized by $\chi^2 = 0.074$ is presented in Fig. 3.3 by dotted lines. It can be seen that the simulation results for the acceptor at positions 7 and 9 clearly show a significant

deviation between simulated and experimental data points. This deviation cannot be explained by small concentration inaccuracies in our sample preparation (Fig. 3.3 A, B). Moreover, the high contribution of positions 7 and 9 to the χ^2 value results in imperfections of the fit for positions 12-18, because the global optimization algorithm tries to decrease the large deviations for positions 7 and 9, rather than to precisely fit all data.

These high deviations lead to the conclusion that a rigid two-helical model cannot describe the protein structure around positions 7 and 9. Therefore it was decided to exclude positions 7 and 9 from the final data analysis and concentrate our research on the data from acceptor positions 12-18. To deal with possible local minima and the stochastic nature of χ^2, the fitting was performed with different initial estimations for 500 times. The best fit is shown in Fig. 3.3 by solid lines. The exclusion of positions 7 and 9 leads to a significant decrease of χ^2: the minimal χ^2 value obtained now is 0.008, which is a factor of 10 smaller than for the previous case.

As in our previous study (Nazarov et al., 2006), we took into account only the best 20% of all solutions found with $\chi^2 \in [0.008, 0.022]$ and discarded solutions with $\chi^2 \in [0.022, 0.203]$. This results in 100 solutions with a good fit to the FRET data (Fig. 3.4 A). Despite the high quality of the fit, a significant uncertainty remains in the angular parameters that describe the tilt and orientation of the N-terminal helix: $\varphi = 15 \pm 13°$, $\xi = -161 \pm 99°$, and $\omega = 61 \pm 62°$. However, the resulting lipid loss parameter $L = 0.20 \pm 0.04$ is quite well defined. To reduce the uncertainty in the angular parameters found, we decided to use additional information coming from the positions of the acceptor fluorescence maxima. Therefore a filtering of the solutions was performed based on the Stokes shift information (Koehorst et al., 2004).

3.4.2. Solution filtering using Stokes shift information

Applying the methodology described in section 3.3.4 to the experimental acceptor fluorescence maxima in Table 3.1 and the resulting protein structures, we were able to filter the solutions by discarding those that do not satisfy the δ criterion (Eq. 3.8). For the resulting 100 solutions, the value of δ varied between 1 and 20. We decided to take into account only solutions with $\delta \leq 2$, which was true for ~50% of the set of solutions (this corresponds to ~10% of the entire set of solutions); the other solutions were discarded. The final set of resulting structures is presented in Fig. 3.4 B. In this figure the two α-helical domains of the protein are shown with solid lines and the "unstructured" region between amino acid residues 1-9 is indicated as a gray cloud.

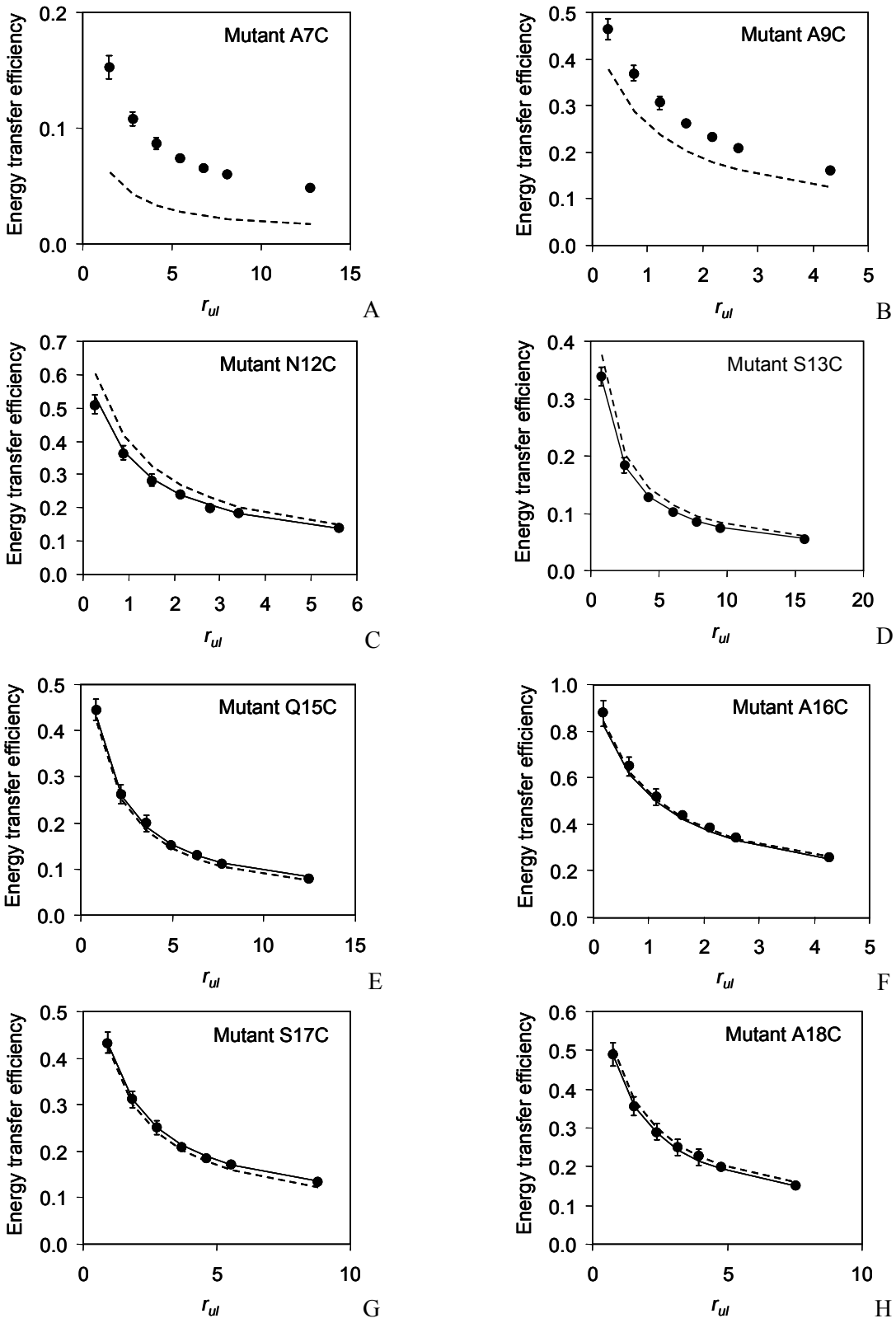


Figure 3.3. Experimental energy transfer efficiencies *E* (filled dots) and their approximation by the model (dotted and solid lines) after global analysis versus the ratio between unlabeled and labeled proteins r_{ul}. The mutant names are given in the right top corner of each plot. Dotted line corresponds to initial fit of data for acceptor label positions 7-18. Solid line presents efficiencies obtained after fitting data for acceptor label positions 12-18.

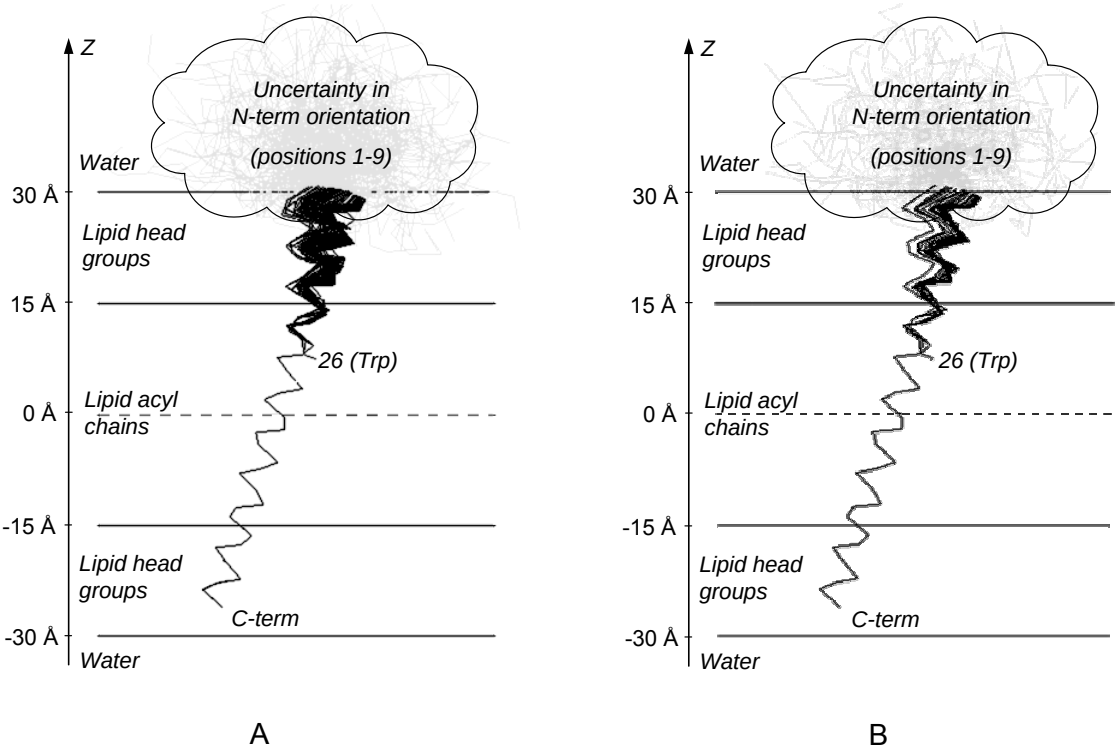


Figure 3.4. (A) Resulting 100 structures obtained from global analysis of experimental FRET data of M13 coat protein in DOPC:DOPG vesicles. The structures are presented in terms of C_α positions that are projected on the plane formed by the OZ axis and the direction of tilt of the transmembrane domain. The protein domain from amino acid residue 1 to 9 cannot be described by a rigid α-helix and is presented as a "cloud" containing several gray "unstructured" conformations. (B) Final set of 52 structures obtained after fitting of experimental data and filtering using Stokes shift information. The resulting tilt angle of the N-terminal domain $\varphi = 5.0 \pm 4.7°$.

The final set of resulting structures indicates a titled I-shaped protein. A kink is determined at position $n_k = 20 \pm 2$, however, the tilt angle of the two protein domains between amino acid residue position 10 and n_k is small: $\varphi = 5.0 \pm 4.7°$. The filtering of the solutions also results in a strong decrease in the uncertainty of the other angular parameters: $\xi = -140 \pm 43°$ and $\omega = 42.1 \pm 10°$. This result indicates that there is a small tilt φ of the N-terminal helix with respect to the transmembrane domain. For such a small tilt, the N-terminal helix tilt direction ξ is not a sensitive parameter, since it describes a small wobble of the N-terminal domain with respect to the transmembrane domain. For an ideal continuous α-helix from the transmembrane to the N-terminal domain ω would be 0°. The resulting value of ω indicates a relatively small distortion of an overall helix at the kink position.

3.5. Discussion

Despite intensive studies, the structure of the membrane-bound state of the M13 major coat protein is still unknown. This is largely due to the difficulty in determining the structure of the N-terminal protein domain. In the literature all sorts of structures are proposed: I-shape (Vos et al., 2005), L-shape (Marassi and Opella, 2003), dynamic (Papavoine et al., 1998; Papavoine et al., 1997), and banana-shape (Spruijt et al., 2000). One of the possible causes of such diversity is the difference in lipid environments. For example, in the solid-state NMR study of Marassi et al. (Marassi and Opella, 2003) the proteins were inserted into dehydrated lipid bilayers. This can lead to squeezing of the proteins and may result in L-shape structures (Vos et al., 2005). Therefore, it is not surprising that in the literature there is not a consistent view about the orientation and tilt of the N-terminal protein domain. It could even be that this domain has no rigid structure and dynamically exchanges between several conformations (Meijer et al., 2001b; Stopar et al., 2002). In our present study, we aimed at minimizing possible artifacts coming from unnatural environments (dehydrated bilayers, micelles) by working at relatively low protein concentrations (high lipid-to-protein ratios) in large unilamellar vesicles. Under such conditions, the application of FRET is ideal, since the technique has a high sensitivity. In the present work, FRET is especially aimed at the problem of the determination of the structure of the N-terminal domain, by taking AEDANS(acceptor)-labeled cysteine mutants in this protein domain and using the natural Trp26 as a donor. To analyze the FRET data, we extended our previous single helix model describing the transmembrane domain of M13 coat protein (Nazarov et al., 2006) to a model of two helical domains that are connected by a helix kink, i.e. the position of the amino acid residue from which the N-terminal helix starts. Furthermore, we took into account the polarity Stokes shift of the AEDANS fluorescence maxima is taken into account by the application of "fuzzy rules" in our data analysis.

The N-terminal protein domain is dominated by the presence of negatively charged amino acid residues (Glu2, Asp4, and Asp5), which will always try to extend into the aqueous phase and therefore act as a hydrophilic anchor (Stopar et al., 2006a). Furthermore, there is a Pro at position 6 (a helix breaker). Therefore, we limited our study to a range of site-directed AEDANS labels attached to the protein from positions 7 to 18. In this range, we decided to leave out positions 11 and 14, since in previous work it was found that these AEDANS-labeled mutants showed an anomalous behavior in the analysis of the fluorescence maximum (Spruijt et al., 2000). Taking into account the yield, quality and availability of mutants, this resulted in eight labeled positions: 7, 9, 12, 13, and 15-18. To discriminate between

intramolecular energy transfer of acceptor-labeled proteins and intermolecular energy transfer, a titration with wild type proteins was performed (Nazarov et al., 2006). Intramolecular energy transfer efficiency is mainly sensitive to the distance between Trp26 and the AEDANS label in one protein molecule, whereas intermolecular efficiency is related to distances between planes, in which donors and acceptors are distributed in the membrane-protein system. In the case of our protein-lipid system, it is not possible to get the protein structure only from intramolecular FRET, since this turns out to be an ill-defined problem. This comes from the fact that we have only a single donor position (the Trp26). The result would not be a single structure, but an infinite number of structures with equal intramolecular distances.

The structure of the protein is studied using a simulation-based fitting approach, which means an adjustment of all variable parameters of the model to fit the simulated data to experimental ones. In this fitting analysis, the parameters that describe the transmembrane helix are taken from our previous FRET study (Nazarov et al., 2006), whereas only the parameters describing the kink position, tilt and orientation of the N-terminal helix domain are varied in our simulations (i.e. n_k, φ, ω, ξ, and L). From the spatial model of the membrane-protein system the coordinates of donors and acceptors are obtained and used to calculate energy transfer efficiencies. To make the analysis more stable we used a global analysis approach, and fit all the data points using the same model (changing only experimental conditions, such as acceptor position n_A, and concentration-dependent ratios r_{LP} and r_{ul}). A validation of our approach is given in Appendix A, where several numerical tests are described and analyzed to determine the precision of the parameters determined. The results indicate that the method can easily distinguish between I- and L-shape protein structures and allows a precise determination of L, n_k, and φ.

Interestingly, in the global analysis of the complete experimental data set, it is found that in our DOPC:DOPG vesicles positions 7 and 9 show a large deviation, indicating that these positions do not fit to the two-helical model. This is consistent with a recent site-directed spin labeling study of M13 coat protein in phospholipid bilayers with increasing acyl chain length (Stopar et al., 2006b). In this work it is found that the N-terminal domain contains 7 unstructured amino acid residues in 22:1PC and 14 residues in 14:1PC. Therefore, it is reasonable to assume that position 7 and 9 are in a flexible or unstructured part of the N-terminal protein domain, for which the rigid helix model does not apply. Consequently, our final analysis was based on a global analysis, excluding positions 7 and 9. The exclusion of

these positions results in a dramatical reduction of the value χ^2, suggesting that for the remaining amino acid residues M13 coat protein is well described by a rigid two-helix model.

To decrease the uncertainty in the angular parameters φ, ω, and ξ we used an additional filtering criterion δ, based on the polarity shift of the AEDANS fluorescence maxima. By applying the "fuzzy rules" polarity criterion given in Eq. 3, we assume that from two AEDANS labels the furthest to the bilayer center is one that has a more red shifted fluorescence (larger Stokes shift). The application of this criterion allows us to discard roughly a half of the solutions, and to more precisely determine the average tilt angle of the N-terminus. This can be seen by comparing Fig. 3.4 A and B, where the two α-helical domains of the protein are indicated with solid lines and the proposed "unstructured" region between amino acid residues 1-9 is drawn as a gray cloud. The best fitting (in terms of both – FRET and Stokes shift) structures for M13 coat protein embedded in DOPC:DOPG vesicles are collected in Fig. 3.4B. Overall, the protein is in a tilted α-helical state from positions 12 to 46 (i.e. the labeled mutants that we investigated here and in (Nazarov et al., 2006)). There is a small kink around position 20, which could indicate that the protein has a weak region in the helix here.

In summary, it can be concluded that the membrane-bound state of the M13 coat protein, showing an overall α-helical conformation, is close to the protein conformation in the intact phage. Such a conformation can be expected to enable a fast and efficient incorporation during the membrane-bound phage assembly of M13 bacteriophage. Probably the overall tilt of the protein is related to an efficient anchoring and integration of the protein in the membrane (Stopar et al., 2006a). Now the structure of the coat protein in a membrane becomes evident, future questions about the membrane-bound phage assembly should address the dissociation of the coat protein from the membrane, i.e. studying the process of lifting the membrane anchors (Stopar et al., 2006a).

We are currently working on further enhancements of our model. One approach that is based on recent findings of Vos et al. (Vos et al., 2005), is to implement the entire AEDANS conformational space for each mutant instead of assuming a constant acceptor arm normal to the helix axis. A further improvement of the precision of our model can be achieved by using all fluorescence data of the AEDANS probe in a general global optimization algorithm. The methodology developed here is not limited to the structure determination of M13 major coat protein and can be used in principle to study the bilayer embedment and structure of any protein (or peptide) for which a one or two-helix model can be applied (Sparr et al., 2005), and with some adaptations to larger membrane proteins.

Acknowledgments. This work was supported by contract no. QLG-CT-2000-01801 of the European Commission (MIVase – New Therapeutic Approaches to Osteoporosis: targeting the osteoclast V-ATPase). We would like to thank Ruud B. Spruijt for the preparation of the protein mutants and helpful comments on the work.

3.6. Appendix A. Sensitivity of the model parameters and noise stability

To determine the sensitivity to the model parameters and the noise stability the following procedure was employed. For each of the two published structures of M13 major coat protein, I-shape (Vos et al., 2005) and L-shape (Marassi and Opella, 2003), artificial FRET data were generated by our model and then used instead of experimental data in the simulation-based fitting algorithm. Because of the stochastic behavior of the χ^2 function the fitting algorithm provides a distribution of solutions for the global minima. The spread of a parameter in this cluster of solutions allows characterizing its sensitivity. To study the experimental noise effects on the parameter distribution, the same operation was performed on data containing artificial noise, similar as is described in our previous work (Nazarov et al., 2006). The standard deviation of the noise varies for each data point (see error bars in Fig. 3.3).

The results of the numerical tests are given in Table 3.3. For an ideal α-helix the algorithm was able to determine the precise structure for the considered range of amino acid residues (12 to 26).

For all solutions in the "elite" set (20% of solutions with smallest χ^2) n_k<12, which means that an ideal helix was found for positions 12 to 26. The introduction of noise to the artificial data did not change this tendency. For an L-shape protein structure the parameters L, n_k, and φ were determined quite well, although the noise in the artificial data increased the uncertainty for almost all parameters. The angular parameters ω and, especially ξ, showed a rather high spread. However, the mean values of the parameters found still were close to the initial values.

Table 3.3. Original and calculated values of the model parameters after analysis of synthetic FRET data by means of a simulation-based fitting approach.

Parameter	Original value in synthetic data simulation	Value found after analysis with no noise added to synthetic data	Value found after analysis with additional noise in synthetic data
I-shape protein (ideal α-helix between positions 12 to 26)			
L	0.2	0.20 ± 0.01	0.19 ± 0.02
n_k	< 12	< 12	< 12
φ	0°	0°	0°
ω	0°	0°	0°
ξ	0°	0°	0°
L-shape protein (Marassi and Opella, 2003)			
L	0.2	0.21 ± 0.02	0.19 ± 0.04
n_k	20	20 ± 1	20 ± 1
φ	110°	107 ± 9°	104 ± 14°
ω	40°	48 ± 12°	53 ± 21°
ξ	-110°	-115 ± 17°	-120 ± 37°

4. ARTIFICIAL NEURAL NETWORK MODIFICATION OF SIMULATION-BASED FITTING: APPLICATION TO A PROTEIN-LIPID SYSTEM

Petr V. Nazarov, Vladimir V. Apanasovich, Vladimir M. Lutkovski,
Mikalai M. Yatskou, Rob B. M. Koehorst, Marcus A. Hemminga

Published in *Journal of Chemical Information and Computer Sciences*
(Journal of Chemical Information and Modeling), **2004**, 44, p. 568-574

ABSTRACT

Simulation-based fitting has been applied to data analysis and parameter determination of complex experimental systems in many areas of chemistry and biophysics. However, this method is limited because of the time costs of the calculations. In this paper it is proposed to approximate and substitute a simulation model by an artificial neural network during the fitting procedure. Such a substitution significantly speeds up the parameter determination. This approach is tested on a model of fluorescence resonance energy transfer (FRET) within a system of site-directed fluorescence labeled M13 major coat protein mutants incorporated into a lipid bilayer. It is demonstrated that in our case the application of a trained artificial neural network for the substitution of the simulation model results in a significant gain in computing time by a factor of $5x10^4$. Moreover, an artificial neural network produces a smooth approximation of the noisy results of a stochastic simulation.

4.1. Introduction

Simulation-based fitting (SBF) has recently become a standard tool for the analysis of experimental data to extract the real parameters of, for example, chemical (Mendes and Kell, 1998; Yatskou et al., 2001a; Yatskou et al., 2001b) and biophysical systems (Berney and Danuser, 2003; Frederix et al., 2002). The idea of SBF is the approximation of experimental data by synthetic data obtained *via* simulation modeling. In comparison to standard analytical data fitting techniques, SBF has the advantage that it fits natural physical and chemical parameters of the system itself and gives a direct insight in how they affect the experimental characteristics of the system (Yatskou et al., 2001b).

However in practice SBF has several limitations. The most crucial problem is that simulation modeling usually is a very time-consuming operation, which results in a long fitting time. In some cases this approach is not useful at all, because the time of optimization becomes non-realistic (from months to years). The origin of another weak point often lies in the stochastic nature of both simulation and experimental data. This results in a very complex behavior of the discrepancy function and the introduction of a large number of local minima (Apanasovich et al., 2000).

The aim of our study is to develop and present solutions for these problems. Here we propose to use an artificial neural network (Bishop, 1995; Wasserman, 1989) (ANN) to speed up the parameter identification and to make the process of fitting less stochastic. The main idea of the method is the substitution of a simulation model by an ANN (specifically a multi-layer perceptron (Bishop, 1995; Stegemann and Buenfeld, 1999; Wasserman, 1989)) during fitting. Because of the simplicity of the multi-layer perceptron structure and the internal mathematics, the computation time needed for the calculation of neural network outputs is much less than for simulation modeling. Hence, the replacement of a simulation model by a multi-layer perceptron leads to a considerable speeding up of all calculations.

The proposed approach of the neural network approximation was tested on a simulation model of resonance energy transfer (Lakowicz, 1999) between fluorescent labels of bacteriophage M13 major coat proteins incorporated into a lipid bilayer.

4.2. Theory

4.2.1. Principles of SBF

The SBF approach was developed for the determination of physical and chemical parameters of complex systems, which cannot be completely described by analytical expressions. Let us consider the idea of SBF on the following general example. A complex (physical or chemical) system Θ can be characterized by a vector of parameters $P = (p_1, p_2, p_3, \ldots)$. These parameters can be regarded as input parameters of the system Θ. After a number of experimental studies on the system Θ are carried out with different input parameters, the vector of output values F can be obtained. In this case, the system can be considered as an operator performing the following operation:

$$\Theta(p_1, p_2, p_3, \ldots) = \Theta(P) = F \,. \tag{4.1}$$

Usually, some input parameters are known. Let us denote them P_0, for example let $P_0 = (p_1, p_2)$. Other parameters, which should be extracted, are denoted as P_X, suppose $P_X = (p_3, p_4, \ldots)$. The vector of input parameters therefore includes a combination of known and unknown parameters $P = (p_1, p_2, p_3, \ldots) = (P_0, P_X)$. The extraction of P_X is the aim of the analysis.

Let us assume that for system Θ it is possible to build an adequate simulation model, which performs operation (4.2) with the same physical parameters P.

$$\Xi(P) \equiv \Xi(P_0, P_X) = F^*, \tag{4.2}$$

where F^* contains the simulated output values, which should approximate the experimental ones.

The determination of the unknown parameters P_X is carried out in the form of SBF. The flow diagram of this method is shown in Fig. 4.1.

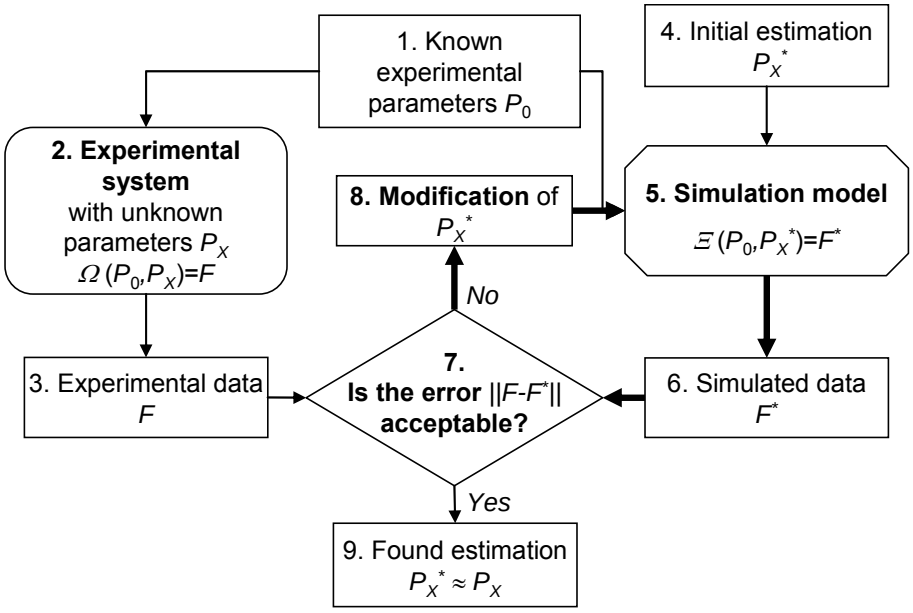


Figure 4.1. Flow diagram, demonstrating simulation-based fitting approach (see text). The thick line shows the fitting loop.

The following steps can be identified in SBF:

1. Output values F are obtained experimentally (see blocks 1-3).
2. An adequate model Ξ of system Θ, which performs operation (4.2), is created (block 5).
3. An initial estimation P_X^* is made for P_X (block 4).

4. An optimization algorithm, using a variation of parameters P_X^*, minimizes the discrepancy function $||F^*–F||$ (blocks 6-8 and 5 again).
5. Finally, the fitted parameters P_X^*, which should estimate the experimental parameters P_X, are obtained (block 9).

As was mentioned before, the main problem of SBF is its time expenses. We solve this problem by the application of an ANN, which approximates and substitutes the simulation model during SBF.

4.2.2. ANN approximation

As was shown independently by Cybenko (Cybenko, 1989) and Hornik (Hornik et al., 1989), continuous smooth functions can be uniformly well approximated by linear combinations and superpositions of sigmoid functions, i.e. by a multi-layer perceptron. This is the most common class of ANNs (Cybenko, 1989; Hornik et al., 1989; Stegemann and Buenfeld, 1999; Tetko et al., 1995; Wasserman, 1989). Concerning the application of ANNs, three layer perceptrons have better learning abilities than two layered ones (Wasserman, 1989).

Most relationships in chemistry and physics can be represented by continuous functions (if they have a stochastic nature – let us speak about their mean). This gives the possibility to approximate the simulation model Ξ by a multi-layer perceptron. Let us denote this approximating ANN transform as Ψ. It performs the operation

$$\Psi(P_0, P_X) = F^{**}, \tag{4.3}$$

where F^{**} is the neural network approximation of the output values of the system.

Hence, instead of the simulation model Ξ, the neural network approximation Ψ can be used during parameter fitting. The suggested ANN approach to parameter determination is illustrated in Fig. 4.2. In this case, the ANN operates as a black-box model of system Θ.

In this approach the following steps can be identified:

1. Output values F are obtained experimentally (see blocks 1-3).
2. An adequate model Ξ of system Θ, which performs operation (4.2), is created (block 5).
3. A representative set of points $\{P\}$ in the parametric space is generated (block 10) and the corresponding simulation values are calculated $\{F^*\}$. This sets form the training set $\{P, F^*\}$ (block 11).
4. The ANN is trained (block 12).

5. An initial estimation P_X^* is made for P_X (block 4).
6. An optimization algorithm, using a variation of parameters P_X^*, minimizes the discrepancy function $||F^{**}–F||$ (blocks 6-8 and 13).
7. Finally, the fitted parameters $P_X^* \approx P_X$ are obtained (block 9).

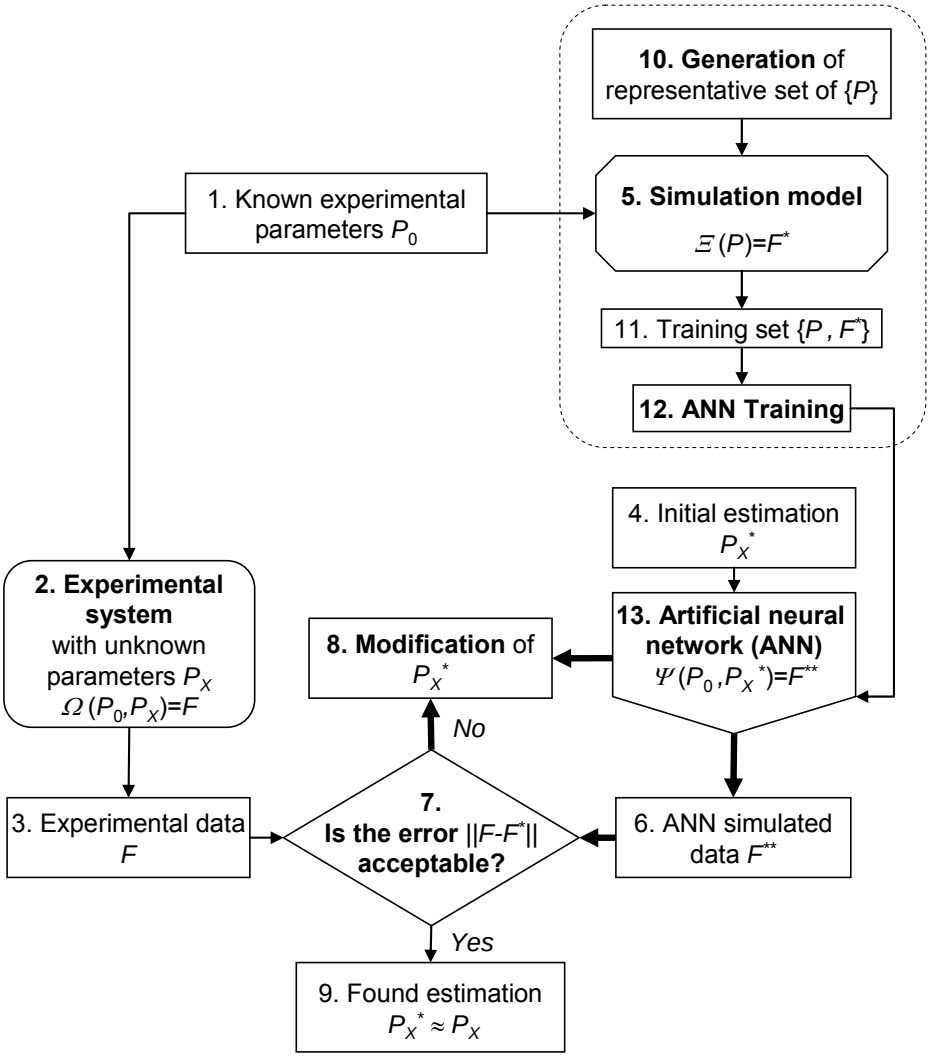


Figure 4.2. Flow diagram of simulation-based fitting with artificial neural network approximation. The dotted box shows the extension of the simulation-based fitting method (see Fig. 4.1) by the artificial neural network fitting procedure.

4.3. Computational

4.3.1. Optimal selection of parameters for the training set

Being replaced by an approximating ANN, the simulation model is used only for initial training of the ANN. The step of the generation of the ANN training set now becomes the most time-consuming part in the proposed scheme in Fig. 4.2, because to obtain each

element of this set the relatively slow simulation should be executed. Furthermore it is of crucial importance for a good approximation to have a representative training set. Therefore to increase the efficiency of a training set, it is necessary to use an algorithm to generate a representative set of parameter points (each point corresponds to a single vector of parameters P describing the system), which are maximally spread in the multidimensional parameter space together with a minimal number of points. Furthermore, to be most flexible, the algorithm should make it possible to increase the number of points without any penalties.

In the present work, the following scheme was developed and applied for the selection of points. It is assumed that each parameter is normalized to the range [0, 1].

1. A set of "boundary" points was generated. For every parameter three values were taken: minimal, maximal and mean. Then all their combinations where taken into account.
2. Main training set generation. Here points are chosen by the following algorithm
    a. Let n be the dimensionality of the parametric space, and N the number of found points. The constant a =1 is preset.
    b. A point with random parametric coordinates is taken.
    c. The distance d from the new point to all previously generated points is calculated.
    d. If the following condition is true

$$\min(d) > \frac{a}{\sqrt[n]{N+1}}, \tag{4.4}$$

    the point is accepted and N is increased by 1. Else, the algorithm checks how many unsuccessful attempts were made before, and if there was a sufficient number of such attempts (in our experiments – 1000), the value of a is decreased by 10%.
    e. The stopping criterion is checked. If it is false, the algorithm goes to step 2.b.

To illustrate the scheme, a space of two-parameters was taken (n=2). The resulting points in comparison with randomly selected ones are shown in Fig. 4.3. Obviously, the application of the scheme allows a uniform infill of the two-dimensional parameter space. The infill itself remains random and can easily be continued. Furthermore, the application of such an approach to the generation of points gives the possibility to select the most distant points of the control set during ANN training (see section 4.3.3) and allows to avoid the selection of points with equal "parametric coordinates". This is important for generalizing the ANN.

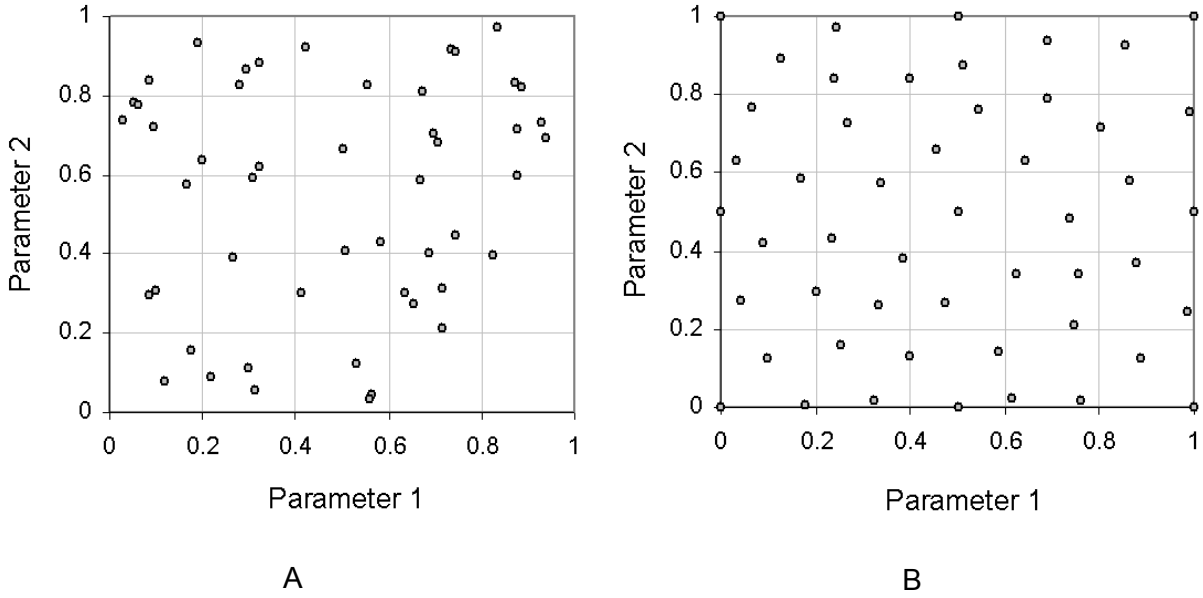


Figure 4.3. Illustration of the principle of point selection in a two-dimensional parameter space. Uniformly distributed random points are shown in (A). Points obtained by the proposed algorithm are presented in (B).

4.3.2. ANN structure

In our research the optimal number of neurons was estimated using the exhaustive search method. In this method, the number of neurons in the first and second hidden layers were optimized. For each number of neurons the ANN was trained for a fixed number of iterations and the resulting training error was calculated. To obtain a statistically valid value for the training error, the training was repeated independently for several times. After all possible combinations of neuron numbers within the region of search, the one with the lowest training error was taken as optimal. The optimal number of neurons found depends on the complexity of the model. These numbers are given in Table 4.2 (see section 4.5).

4.3.3. Training of the ANN

Upon training of the multi-layer perceptron two rather contradictory conditions should be satisfied. From the one side the mismatch between desired and obtained outputs should be decreased. From the other side an ANN should not lose its generalization abilities (Stegemann and Buenfeld, 1999). An excessively long training results in a sliding of the ANN coefficients to a local minimum. This makes the approximation worse in points that do not belong to a training set. This phenomenon is called overtraining (Stegemann and Buenfeld, 1999; Tetko et al., 1995). Therefore, a special training strategy, based on the generation of an additional

small control set, was applied to deal with this problem. After each epoch of training the performance of the ANN is verified on this control set. The training is terminated if the performance does not change or decrease for a certain number of epochs (Tetko et al., 1995). Numerical calculations were performed to determine the optimal size of the control set. The training set was separated into an actual training set and a control set with the following ratios: 90%-10%, 85%-15%, 80%-20%, 75%-25%, and 70%-30%. For the 3-parametric model the best performance was obtained for the 80%-20% ratio, for the 4-parametric model this was 75%-15%, and for the 5- and 6-parametric model the ratio 90%-10% gave the best results. The application of the proposed method of training set generation (section 4.3.1) allows a simple separation procedure: the last generated elements of the training set should be taken for the control set.

In our calculations the ANN is trained by the back propagation error algorithm with the Levenberg-Marquardt optimization technique (Hagan and Menhaj, 1994).

4.4. Experimental objects and methods

The proposed approach of the neural network approximation, as shown in Fig 4.2, was tested on a simulation model of fluorescence resonance energy transfer (FRET) between fluorescent labels of bacteriophage M13 major coat protein mutants incorporated into a lipid bilayer.

4.4.1. FRET

The idea of FRET spectroscopy is based on a dipole-dipole radiationless energy transfer and was initially developed by Förster (Förster, 1948) and further enhanced by Stryer (Stryer, 1978). Macromolecules studied (in our case – membrane proteins) are labeled with fluorescent probes of two types: donors and acceptors (Lakowicz, 1999). The emission spectrum of the donor and the absorption spectrum of the acceptor should overlap. Donors are excited by an external light source and some of them transfer excitation energy to acceptors due to dipole-dipole radiationless energy transfer. The probability of energy transfer for an isolated donor-acceptor pair is:

$$p_{ET} = \frac{1}{1 + \left(r/R_0\right)^6} \ , \tag{4.5}$$

where r is the distance between the donor and the acceptor, and R_0 is the so-called Förster distance, which corresponds to 50% energy transfer probability *via* dipole-dipole interaction (Förster, 1948). For a system containing n_A acceptors, the expression for the energy transfer becomes somewhat more complex:

$$p_{ET} = \frac{\sum_{i=1}^{n_A} (R_0/r_i)^6}{1 + \sum_{i=1}^{n_A} (R_0/r_i)^6}. \tag{4.6}$$

The mean probability of energy transfer in the system, containing n_D donors and n_A acceptors, is called the energy transfer efficiency and can be calculated as a mean value of energy transfer probabilities for all donors:

$$E = \frac{1}{n_D} \sum_{j=1}^{n_D} (p_{ET})_j . \tag{4.7}$$

By observing the energy transfer process one can get information about the relative location of donor and acceptor labels.

4.4.2. Biophysical protein-lipid model

The membrane-bound major coat protein of M13 bacteriophage, which infects *E. coli*, is an excellent model system to study fundamental aspects of protein-lipid and protein-protein interactions. This single membrane-spanning protein consists of 50 amino acid residues and has mainly an α-helical conformation. The protein has been extensively studied in membrane model systems by biophysical techniques (Meijer et al., 2001a; Spruijt et al., 2000; Stopar et al., 2002; Stopar et al., 2003).

For FRET studies, the natural amino acid residue tryptophan of M13 major coat protein at position 26 was used as a donor label. To introduce an acceptor label to the protein, a number of mutants, containing unique cysteine residues at specific positions, was produced. The cysteine residues were specifically labeled with the fluorescent environmental probe N-(iodoacetylaminoethyl)-5-naphthylamine-1-sulfonic acid (AEDANS) (Spruijt et al., 2000). This fluorescent label was used as an acceptor. Since the labeling efficiency with AEDANS is less than 100% the entire protein-lipid system contains proteins of two types: unlabelled proteins – with the natural donor, and labeled ones – with both donor and acceptor.

To study such a complex system the following simplified spatial model was designed. The biological membrane is approximated by a two-dimensional periodic structure with a hexagonal packing of the lipids in which the M13 coat protein mutants are distributed (Fig. 4.4 A). The area occupied by each membrane protein on the membrane surface is assumed to be equal to that of two lipids. It is assumed that the distance between two nearest molecules on the grid is 8.0 Å and the thickness of the lipid bilayer is 30 Å. The α-helical M13 coat protein mutants are approximated by rods with a constant location of the donor (D) and a variable location of the acceptor (A) (see Fig. 4.4 B).

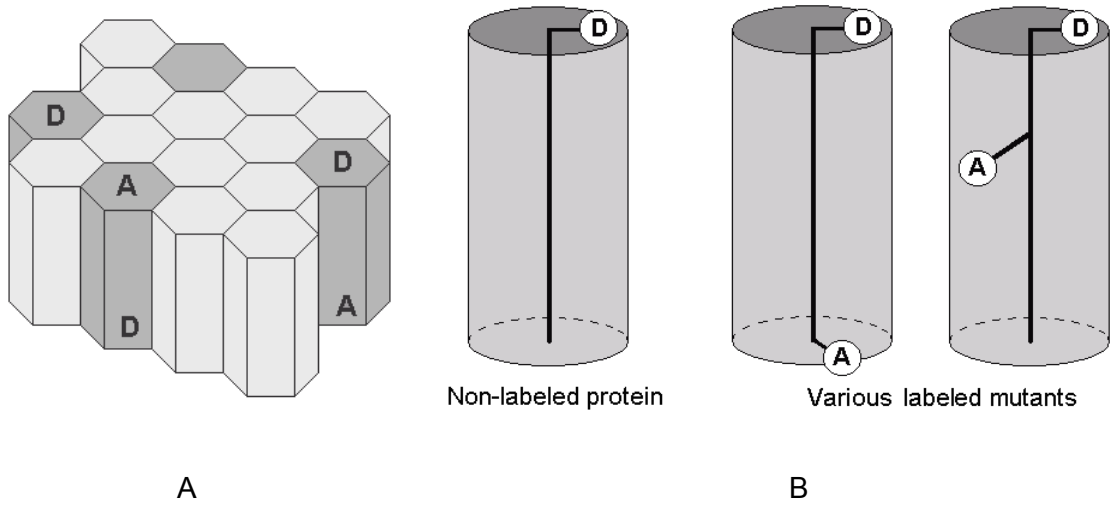


Figure 4.4. Model of a membrane (A) and a membrane protein (B) with fluorescent labels. The non-labeled protein only contains a tryptophan. The protein mutants have acceptor label (AEDANS) at various positions along the protein structure. The protein is assumed to be α-helical.

4.4.3. Simulation of energy transfer

The input parameters of the model are presented in Table 4.1. The ranges listed in this table are selected such that they are physically valid, and cover all possible experimental situations.

In the experimental situation, parameters 1-3 and 6 are known (some of them within a small experimental inaccuracy), thus they can be regarded as P_0 in terms of the parameter description in section 4.2. However, in general the separation of the parameters in Table 4.1 into P_0 and P_X depends on the situation. For instance, for the three-parametric model parameters 1 and 2 can be used as P_X, and parameter 3, which is known precisely, as P_0. It should be mentioned that the three-parametric model was used only for the validation of the

methodology. For the 4-parametric model, the coefficient of protein association becomes the subject of interest (P_X), while parameters 1-3 are the known parameters (P_0).

Table 4.1. Input parameters of the simulation model.

Number	Parameter	Description	Range
1	Surface density of labeled proteins	The ratio of the area occupied by labeled proteins (containing the donor and acceptor) to the area of the entire membrane.	0.0001 ÷ 0.1
2	Surface density of non-labeled proteins	The ratio of the area occupied by non-labeled proteins (containing only donor) to the area of the entire membrane.	0.0001 ÷ 0.1
3	Labeling site	The amino acid residue number to which the acceptor is attached.	1 ÷ 50
4	Coefficient of protein association	The probability that a selected protein is located in the immediate proximity to another one.	0 ÷ 1
5	Size of molecules	The minimal distance between the centers of 2 nearest molecules (proteins and lipids).	5 ÷ 10 Å
6	Förster distance	Donor-acceptor distance (for an isolated pair) corresponding to 50% energy transfer.	1 ÷ 100 Å

The fluorescence intensity and energy transfer efficiency for the entire protein-lipid system are taken as output values (F in terms of the description in section 4.2). Because of the simulation nature of the model, the resulting output contains stochastic errors. Therefore simulations are run several times to reduce these errors. The flow diagram of the simulation is shown in Fig. 4.5.

The simulation is carried out in the following way:

1. The parameters of the system are set (block 1).
2. A spatial model of the membrane with embedded proteins is created in accordance with the input parameters. The coordinates and orientation of the proteins provide information about the locations of donors and acceptors in the system (block 2).
3. For each donor (denoted as i_D) the distances to all acceptor are considered and the probability of energy transfer (to any of them) is calculated using Eq. 4.6 (blocks 3-5).
4. The mean probability of energy transfer among all donors results in the energy transfer efficiency for the whole system.
5. Steps 2-4 (and blocks 2-6 in the flow diagram) are repeated for several times to decrease the effect of the randomness of the protein distribution. In our calculations we used an empirical value of $10^4/n_D$ simulations, where n_D is the number of donors.

Additional simulations using this model and experimental FRET data will be published elsewhere (see Chapter 2 and 3).

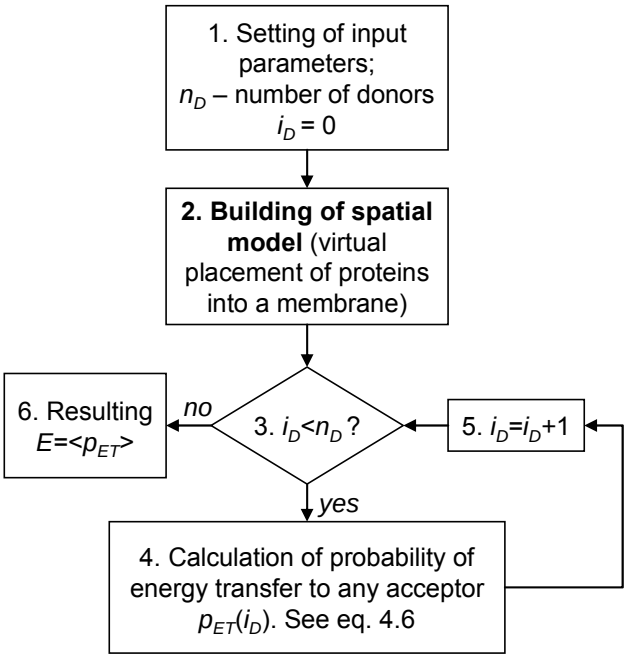


Figure 4.5. Flow diagram of a single simulation of energy transfer in a protein-lipid system.

4.5. Results and discussion

4.5.1. ANN configuration

Before transferring the simulation model to the ANN, the input parameters (Table 4.1) were normalized to the range [0,1] by the simple linear mini-max method. Thus, input values 0 correspond to minimal possible parameter values, and 1 to the maximal ones.

The optimal number of neurons obtained experimentally varied with the number of input parameters. These numbers are presented in Table 4.2.

After the number of neurons in the ANN was determined, it was trained as was described in section 4.3. To avoid overtraining, after each 10 epochs the ANN was tested on a control set. If the result of testing did not improve for 30 epochs, the training procedure was stopped. The resulting mean relative square error on the training set varied up to 2%.

Table 4.2. Optimal number of neurons in the ANN used. The input parameters are described in Table 4.1.

Number of input values	3	4	5	6
Model parameters	1-3	1-4	1-5	1-6
Number of neurons in the first layer	13	15	18	20
Number of neurons in the second layer	10	13	16	20

4.5.2. Time costs

All calculations in this article were made in MATLAB® 6.1 with the Neural Networks Toolbox on a PC with Intel Pentium III-850 CPU. The time costs of the ANN method application in this set up are shown in Table 4.3.

Table 4.3. The time costs of the ANN approximation of the FRET simulation model.

Number of parameters	3	4	5	6
Time for generation of the training set	11 hr	22 hr	56 hr	110 hr
Time for training	6 min	10 min	14 min	20 min
Time for ANN simulation	6.0×10^{-4} s	7.0×10^{-4} s	8.0×10^{-4} s	10^{-3} s
Average time for simulation modeling	40 s	40 s	40 s	40 s
Average gain in computer time	6.7×10^{4}	5.7×10^{4}	5.0×10^{4}	4.0×10^{4}

From Table 4.3, it is clear that the generation of training sets is the most time-consuming operation. However, it should be noted that this process does not need the supervision of a human and the training itself needs to be performed only once for each simulation model. The gain in computing time, which is about $5x10^4$, does not decrease significantly with the increase of the network complexity.

A typical illustration of the ANN approximation of the simulation model is shown in Fig. 4.6. Here the energy transfer efficiency is plotted as a function of the location of the acceptor. The oscillations in this plot arise from the α-helical nature of the protein model as is shown in Fig. 4.4 B. The relative deviation between the ANN approximation and the actual model calculation is less than 3%, showing that the ANN approximation performs very well.

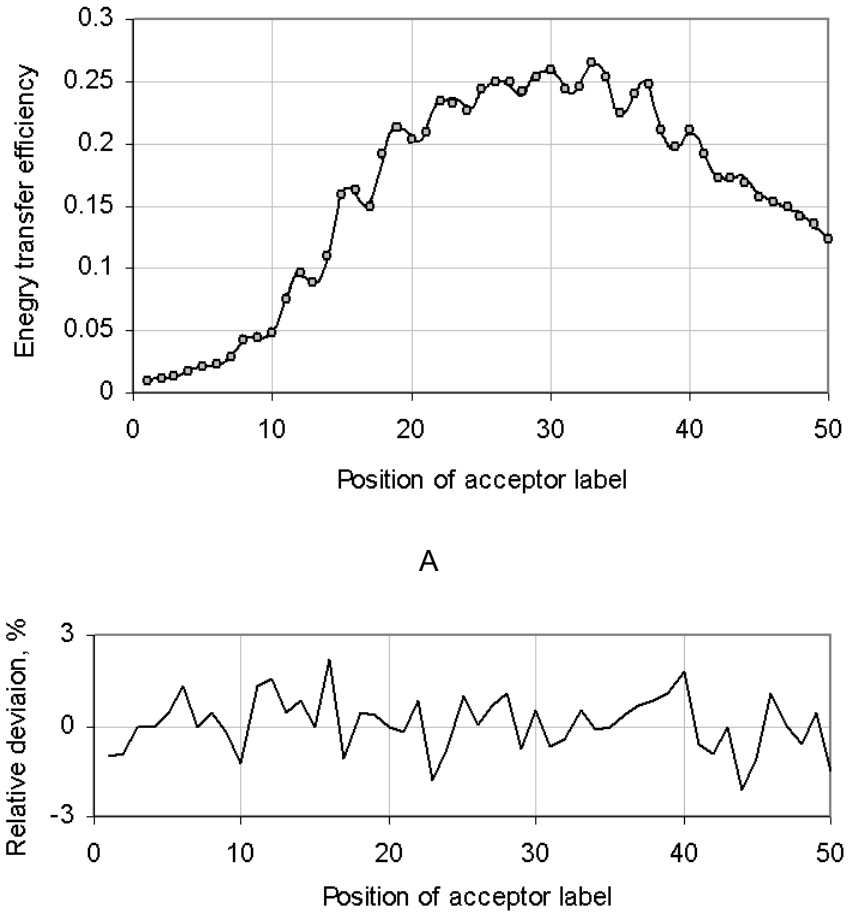


Figure 4.6. ANN approximation of the simulation model (A) and relative deviations of the ANN result and the simulation model (B). In (A) the circles show the result of the simulation modeling and the line is the ANN approximation.

4.5.3. Consistency of the approximation

To obtain information about the consistency of the ANN approximation we conducted several statistical calculations on the 4-parametrical model. In these calculations we modified one of the model parameters – the association coefficient, and analyzed the deviation between the ANN approximation and the simulation modeling. In Fig. 4.7 A the energy transfer efficiency is plotted for various values of the protein association constant. The agreement between the ANN approximation and actual model calculation is good.

Since the auto-correlation function, calculated from the deviations (Fig. 4.7 A), is close to a delta-function, we conclude that the deviations are not correlated and behave stochastically. The distribution of the deviations (Fig. 4.7 B) is close to a Gaussian line shape, indicating that the deviations are the result of the randomness of the simulation model.

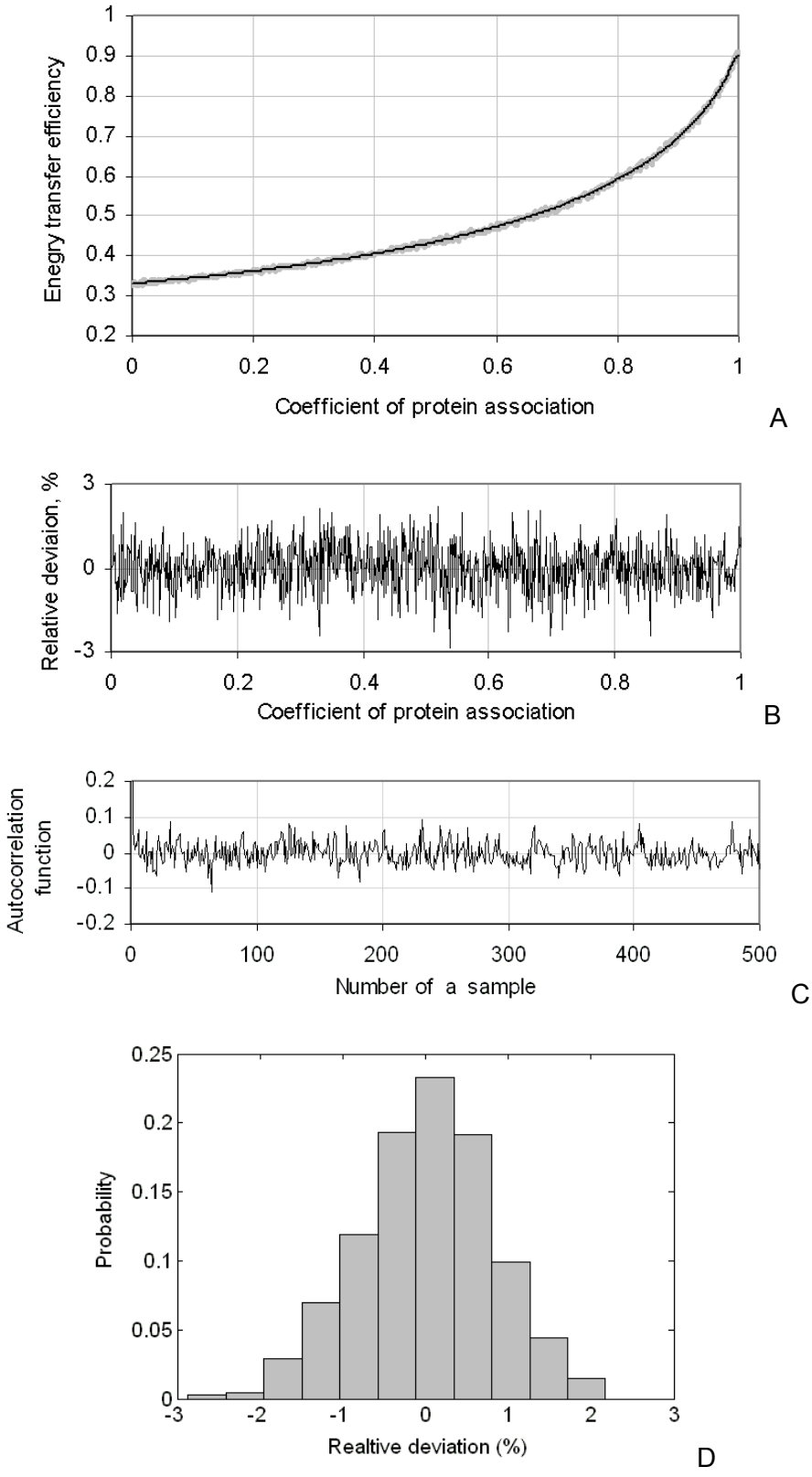


Figure 4.7. Consistency of the ANN approximation. In (A) the thick gray line is the result of the simulation modeling and the thin black line is the ANN approximation. Below this graph the relative deviations between simulation model and approximation (B) and their autocorrelation function (C) are given. The distribution of deviations is given in (D).

It should be mentioned that a multi-layer perceptron with sigmoid activating functions produces a smooth approximation of a stochastic simulation model. This approximation does not contain stochastic noise. Thus, the fitting procedure operates with a less stochastic discrepancy function $||F–F^{**}||$, and therefore contains less local minima than in the case of stochastic simulation model fitting.

4.6. Conclusions

The use of a trained ANN in the biophysical modeling presented here results in a gain in computing time by a factor of 5×10^4. Moreover an ANN produces a smooth approximation of the results of a stochastic simulation. Thus it decreases the level of stochastic errors. Due to this smooth dependency it will simplify the application of standard optimization techniques, such as gradient search, for parameter determination. It was shown that the deviations between the actual model outputs and its ANN approximation have a stochastic nature. In our case the relative deviation was less than 3%.

The approach used in our calculations has some imperfections. It works only when the number of variable parameters is relatively small (in our calculations up to 6). Furthermore the calculations related to the generation of the training set are quite time extensive, although they need to be performed only once for a simulation model.

In conclusion, the method of ANN modeling is an excellent tool for determination of parameters of specific systems. In fact the method can be generalized to analyze any experimental system for which SBF can be applied.

5. NEURAL NETWORK DATA ANALYSIS FOR INTRACAVITY LASER SPECTROSCOPY

Petr V. Nazarov, Vladimir V. Apanasovich, Katsiaryna U. Lutkovskaya,
Vladimir M. Lutkovski, Pulat Y. Misakov

Published in *Proceedings of SPIE*, **2003**, 5135, p. 61–69.

ABSTRACT

The method of data analysis in intracavity laser spectroscopy is considered. The artificial neural network was used as an analyzing tool for the determination of elements concentration in trace amounts samples using absorption spectra. The special neural network training algorithm based on simulation of experimental spectra was developed to solve the problem of non-sufficient experimental data set. The application of this method allows achieve the better sensitivity than conventional analytical methods and proved itself more robust. The proposed method was tested on spectra of Cs water solutions.

5.1. Introduction

Laser techniques are widely used for determination of trace amounts of elements, especially toxic and biologically significant, in environmental samples and human tissues. The intracavity laser spectroscopy method used in this work is one of the most sensitive analytical techniques.

It was shown that advanced data processing based on digital filtering and regressive analysis improves the detection limits (Burakov et al., 2002). The mean-square error for the trace concentrations of cesium in water solutions was decreased to the value less then 10%.

Unfortunately, practically available conventional analytical devices and linear data processing have reached their limits. The relationship between evaluated parameters and the experimental input parameters is nonlinear and very complex. It does not allow to eliminate all errors in experimental data. This is the reason to develop new approaches for data processing in the high-resolution spectroscopy.

Adaptive nonlinear algorithms based on artificial neural networks (ANNs) are regarded in this contest as a very attractive tool. ANN is a data processing system consisting of a large number of simple highly interconnected processing elements. It utilizes the weight matrices to perform the mathematical transformation of the input vector to the output vector.

Special learning procedures are used to adjust the weight matrix for required relationship between input and output of ANN. It is quite different from the traditional computing system. ANNs have some remarkable properties such as flexibility, capability of learning and generalization. Special software tools for emulation of ANNs in conventional computers are available nowadays and this approach is becoming more and more popular (Bishop, 1995).

The brief review of ANN implementation in data processing algorithms is presented in the next section. The intracavity laser spectrometer used as the experimental setup is described in the section 5.3.

Experimental data and preprocessing technique are reviewed in the section 5.4.

The neural network procedure for data processing is presented in the section 5.5. Finally the received measurement errors are estimated and conclusions are made in the last section 5.6.

5.2. Neural networks as a data processing tool

A solution of inverse problems is one of the most important attributes of spectroscopy. It was shown that ANNs allow solving both direct and inverse problems using standard experimental data involved in the calibration procedure. The unique opportunities of ANN were used for solving the following inverse problems (Gedova et al., 2002):

- precise determination of water temperature from Raman spectra,
- determination of small fluorescent contributions for components of an organic compounds mixture in water from their fluorescence spectra,
- determination of molecular parameters of organic compounds from fluorescence spectra,
- time-resolved kinetic spectroscopy performed with long excitation pulse and a detector with low temporal resolution.

There are examples of applications of ANN for signal processing of transient atomic absorption signal (Kale and Voigtman, 1995), classification and recognition of spectra (Eghbaldar et al., 1998; Pulido et al., 1999; Ramadan et al., 2001; Schulz et al., 1995; Walczak, 1996), evaluation of extremely low concentrations of analyzed substances in soil (Schulz et al., 1995) and water samples (Apanasovich et al., 2001; Ruisanchez et al., 1997; Walczak and Massart, 1996a; Walczak and Massart, 1996b).

Neural networks based on radial basis functions (RBF) were applied to the classification of visible and ultraviolet spectra and they were implemented in the auto-diagnosis process of a flow injection analytical system. The classification error was 13%, which was a significant reduction compared with the 20% when using counterpropagation

neural networks as the classification technique. The importance of this reduction lies in the fact that the number of analytical errors which have a considerable effect on the system is reduced to half. The spectra did not have to be preprocessed to distinguish between the five classes and this means less extra work and a reduction in computation time. The training procedure is also simpler in the case of using RBF because there are fewer parameters to optimize. With counterpropagation neural networks the number of epochs, the number of neurons, the type of transfer function and the initialization of the weights have to be optimized but when using RBF only two parameters have to be optimized: the width of the radial functions and the number of neurons in the hidden layer (Pulido et al., 1999). Thus neural networks based on RBF proved to be a useful tool in the classification of ultraviolet and visible spectra.

The same methodology with the backpropagation ANN (or multilayer perceptron, MLP) was used for encoding and pattern recognition of infrared spectra (Schulz et al., 1995).

ANN were also used for determination of Ru in flow-injection analysis systems (Wang et al., 2001) and as catalyst for simultaneous determination V (IV) and Fe (II) through the single catalytic kinetic run (Safavi et al., 2001) due to high speed data processing. Counterpropagation (Ruisanchez et al., 1997; Schulz et al., 1995), backpropagation (Eghbaldar et al., 1998; Walczak, 1996) or RBF (Pulido et al., 1999; Walczak and Massart, 1996a; Walczak and Massart, 1996b) neural networks my be used for achieving better accuracy depending on problem. Thus flexibility and universality of ANNs are the important advantages of the discussed approach.

5.3. Experimental setup

The intracavity laser spectrometer was used in the experiment. It consists of the four basic modules: dye laser, electrothermal atomizer with graphite furnace for liquid samples, high-resolution spectrograph and data processing system as it is shown schematically in the Fig. 5.1.

The tunable dye laser is used as the primary light source. Being flash-lamp pumped it radiates a smooth broad-band spectrum in the range 440 - 700 nm. The spectrum width depends on a dye type and usually is 10 -15 nm. A Fabri-Perot interferometer in laser cavity was used in some cases for dye laser spectrum stabilization near an absorption line. With an interferometer the maximal width of a laser spectrum decreases to 1.0 - 1.5 nm. Laser pulse duration can easily be changed between 1 and 10 μs by variation of power supply parameters. For the measurement of dye laser pulse duration a silicon photodiode is used.

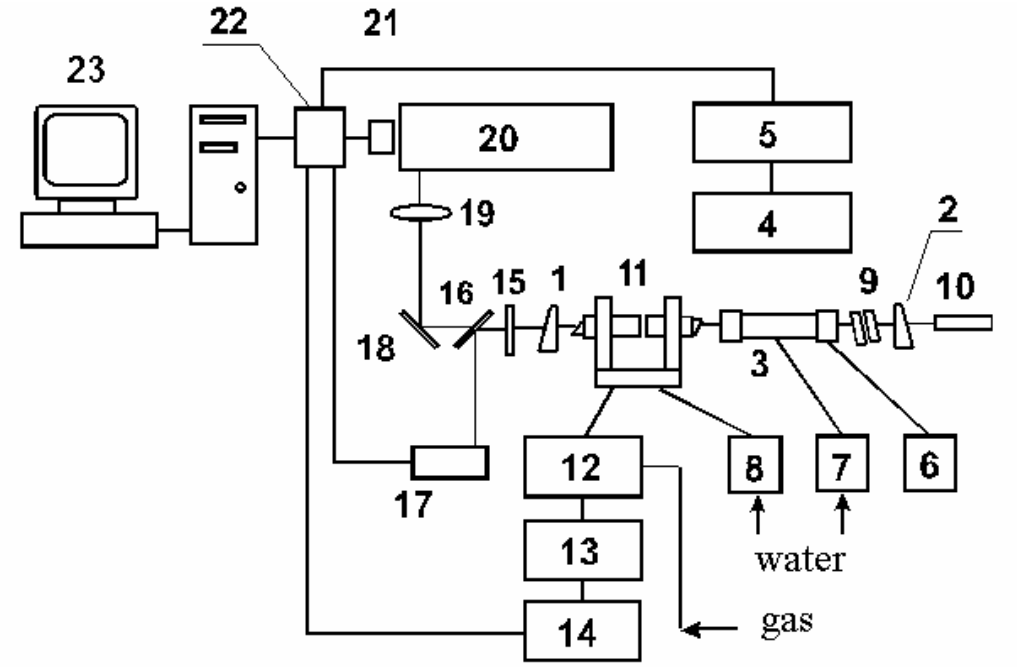


Figure 5.1. Intracavity laser spectrometer: 1 - output cavity mirror, 2 - back cavity mirror, 3 - dye laser, 4 - laser power supply, 5 - laser control unit, 6 - dye pump, 7 - laser water pump, 8 – atomizer water pump, 9 - Fabry-Perot interferometer, 10 - He-Ne laser, 11 - atomizer, 12 - atomizer gas supply, 13 - atomizer power supply, 14 - atomizer control unit, 15 - attenuator, 16 - beam splitter, 17 -photodiode, 18 - mirror, 19 - cylindrical lens, 20 - high-resolution spectrograph, 21 - CCD-camera, 22 - optical multichannel analyzer, 23 –PC.

A graphite furnace, electrothermal atomizer was used in the intracavity laser spectrometer for atomization of cesium samples. The atomizer was located in the laser cavity between dye cell and output mirror. It has special wedge windows for preventing interferometric structure in dye laser spectrum. Spherical cavity mirrors were used for the same purpose as well as for collimating of dye laser radiation inside a cavity through a graphite furnace. Moreover wedge cavity mirrors were employed, but in this case it is desirable to adjust the inner diameters of the furnace and the dye cell. The graphite tube was 28 mm long with a 6 mm inner diameter and 8 mm outer diameter. The atomizer has 20 - 3070° C heating temperature interval and 64 heating steps with the step duration 1 - 799 s.

Stock and standard solutions were prepared by using de-ionized triply distilled water in accordance with a conventional sample preparation procedure (Whiteside, 1988).

Atomic absorption signals were measured at the cesium wavelength of $\lambda = 455.531$ nm. The use of a 5 μs dye laser makes it possible to provide an effective length of absorbing layer of about 100 m for geometrical length of a graphite furnace of 28 mm.

Dye laser spectra with absorption lines were recorded with the help of a 0.001 nm resolution echelle spectrograph with an optical multichannel analyzer. The charge coupled device (CCD) array was used for detection of optical radiation. The grating (300 lines/mm) operating in high orders of the spectrum (6 - 25) with double dispersion was used as a dispersion element. The focal length of the objective was 1377 mm, the relative aperture was 1:21. Processing of obtained data as well as controlling of spectrometer modules was performed by the personal computer (PC).

5.4. Experimental data and its preparation

5.4.1. Absorption spectra

The developed method of spectra analysis was tested on absorption spectra of cesium in water solutions. The Cs I resonant line at λ = 455.531 nm corresponding to the transition 6s $^2S_{1/2}$ - 7p $^2P^0_{3/2}$ was used as analytical one. The typical raw absorption spectrum of the Cs 455.531 nm line is presented in Fig. 5.2.

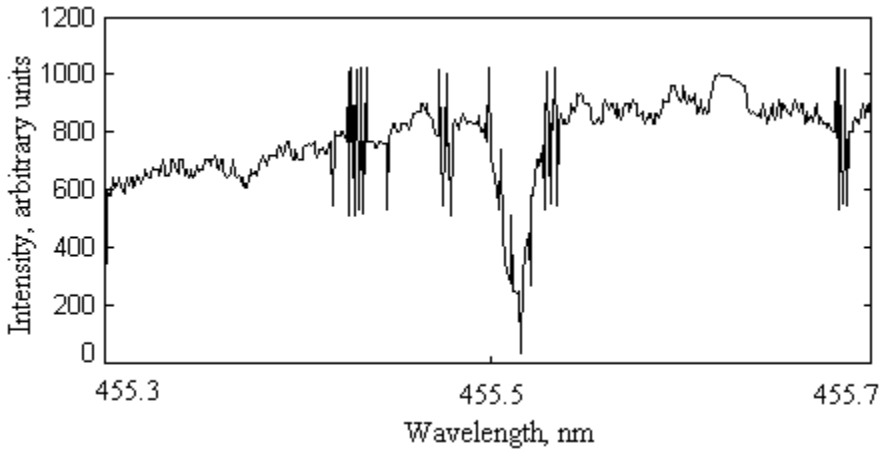


Figure 5.2. Raw absorption spectrum of Cs (narrow gap) on top of the laser pulse spectrum

The most commonly used function of element concentration definition in the intracavity laser spectroscopy is relative depth of dip $\Delta I/I_0$ (Fig. 5.3). In some works equivalent breadth of absorption $\Delta S/I_0$ was considered as a spectra parameter. Physical meaning of that parameter could be seen from the Fig. 5.3. These parameters are used in classical approaches to analyze absorption spectra. Such approaches were realized and the rate mean square errors (RMSE) of those methods come to approximately 10%.

The main factors which make processing of spectrums difficult are:

1. Non-stable time and spectral shape of lasing pulse. This results in the change of the base line and its non-linearity.

2. The presence of interference components in a spectrum. A base line has irreproducible structure, which is similar to a low-frequency noise. This interference spectral structure appears in the cavity of the laser and exerts multiplicative influence upon an absorption dip.

3. A multiplicative high-frequency noise of a CCD detector.

4. A noise appearing in detection system during a laser pulse. High power electrical discharge in flash lamps supply initiated errors in analog-digital converter.

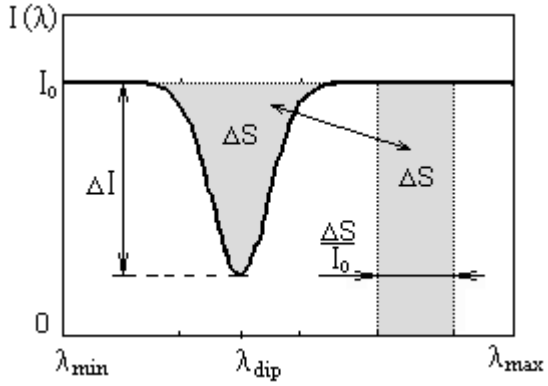


Figure 5.3. Parameters of absorption spectra: relative depth of the gap $\Delta I/I_0$ and area $\Delta S/I_0$

5.4.2. Preprocessing of experimental spectra

The application of preprocessing procedures, aimed to diminish the influence of disturbing factors, is the first step in any processing procedure of real-world signals.

Before analysis a spectrum should be cleared from glitches (sharp over fall noise) because it does not carry any information. To achieve this purpose a spectrum was looked for consecutive samples with highly different values (like δ-peaks). If such a sample was found, its value was changed to the mean of the nearest samples.

The filtration of spectrums was realized in the following way. The Fourier transform of a spectrum was calculated. Its real and imaginary parts were multiplied by a window of the special shape (several functions were tried, and the best results were obtained with the help of Gaussian window). After that the inverse Fourier transform was applied. A result of such filtration is shown in the Fig. 5.4.

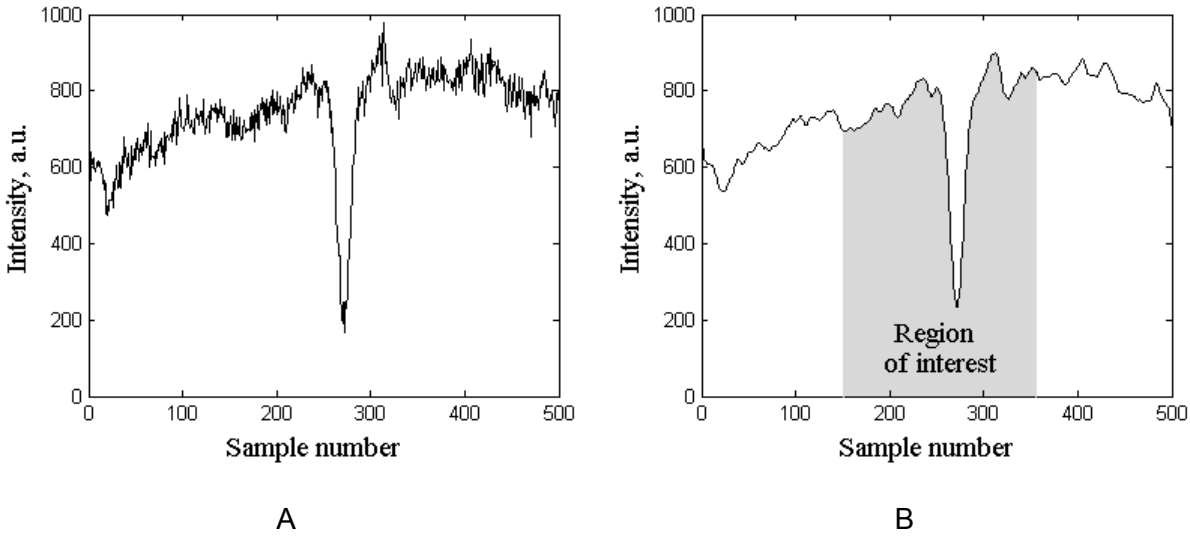


Figure 5.4. High frequency noise cancellation by Fourier filtration: (A) – initial spectrum, (B) – filtered spectrum.

5.5. Neural network processing of absorption spectra

5.5.1. Application of ANN

The main idea in application of ANN for analysis of absorption spectra is the following. The central and most informative part of the preprocessed (see Fig. 5.4 B) and normalized spectrum was used as inputs of neural network. The information is processed and the normalized assumed concentration of element was obtained from an output.

The 3-layer feed-forward neural network (so called multilayer perceptron) was used for analysis of absorption spectra (Fig. 5.5). ANN was trained by gradient descent with momentum and adaptive learning rate backpropagation error method.

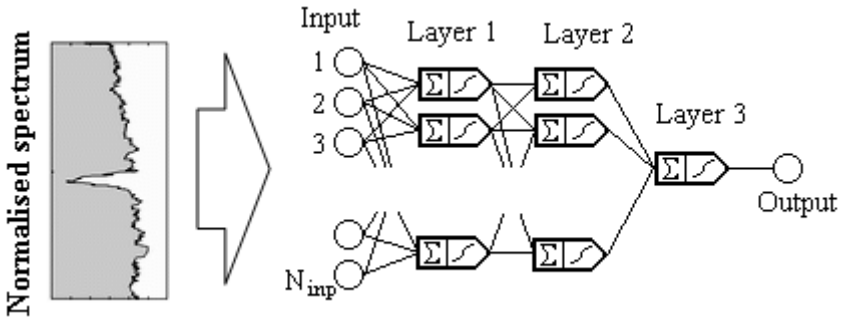


Figure 5.5. Three-layer feed-forward neural network (MLP) for analysis of normalized spectrum.

There are several problems of such application of ANN. These problems and proposed solutions are listed below in sections 5.2 and 5.3.

5.5.2. Avoiding of the lack of experimental training pairs

ANN usually needs quite large representative training set to work in a proper way. Unfortunately it is very difficult to obtain enough training pairs from an experiment because of high time costs and material charges. To avoid this computer simulation of additional training pairs was used as follows. The spectra from primary training set are analyzed and approximated by deterministic and stochastic functions as it is shown in the Fig. 5.6.

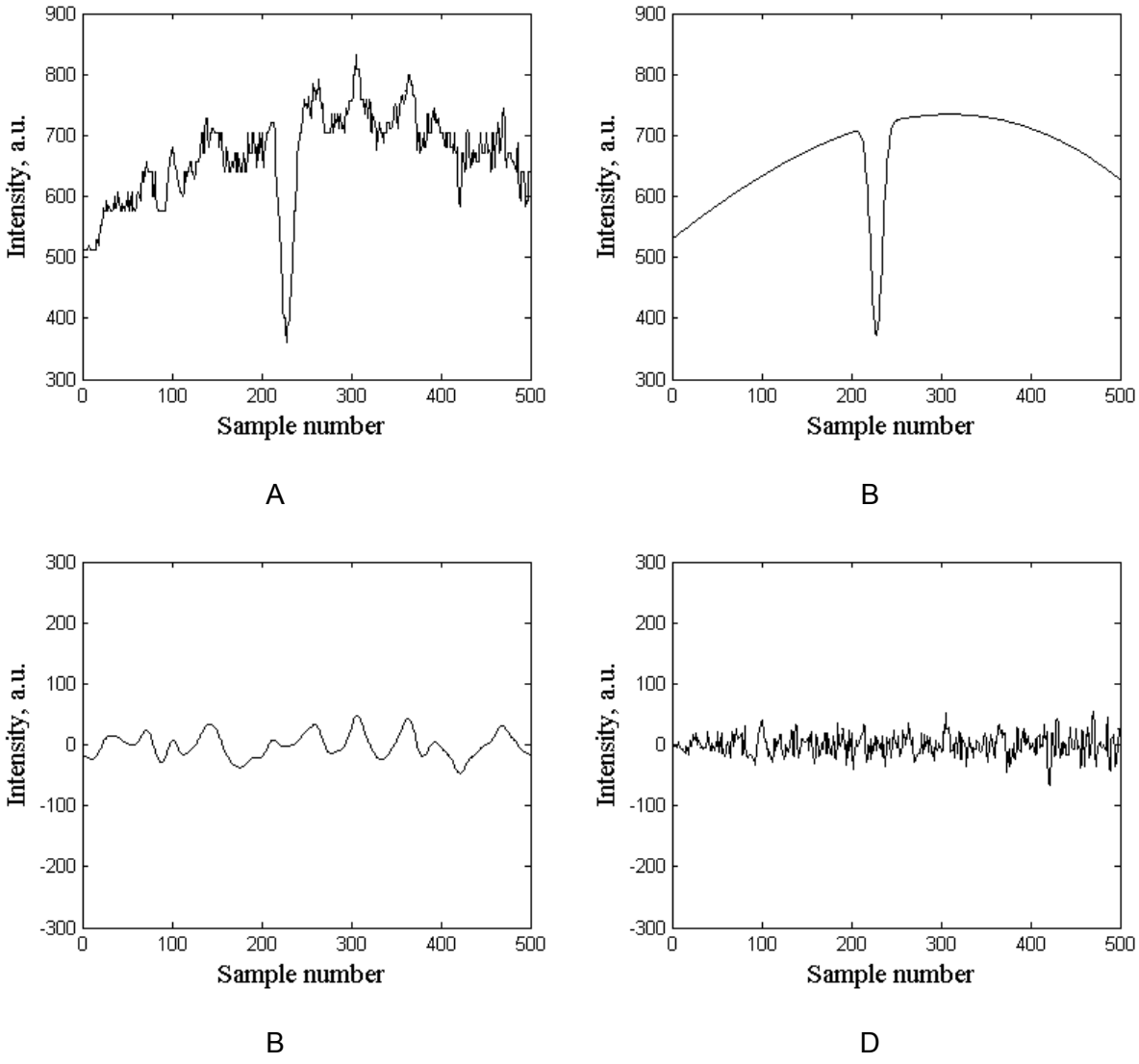


Figure 5.6. Separation of initial spectra to compounds for simulation (A) – initial spectrum, (B) – ideal laser spectrum with absorption peak, (C) – low frequency and (D) – high frequency noises.

Artificial spectra were created to form a new training set for ANN. This simulated training set was used for initial training of a neural network. It is obvious that the performance of ANN trained in such a way is quite rough. To make it more exact ANN is trained once more on experimental training set. This algorithm is shown in Fig. 5.7. The robustness of the algorithm to the overtraining problem was solved by checking the ANN efficiency on the control training set, as it was proposed in (Gedova et al., 2002).

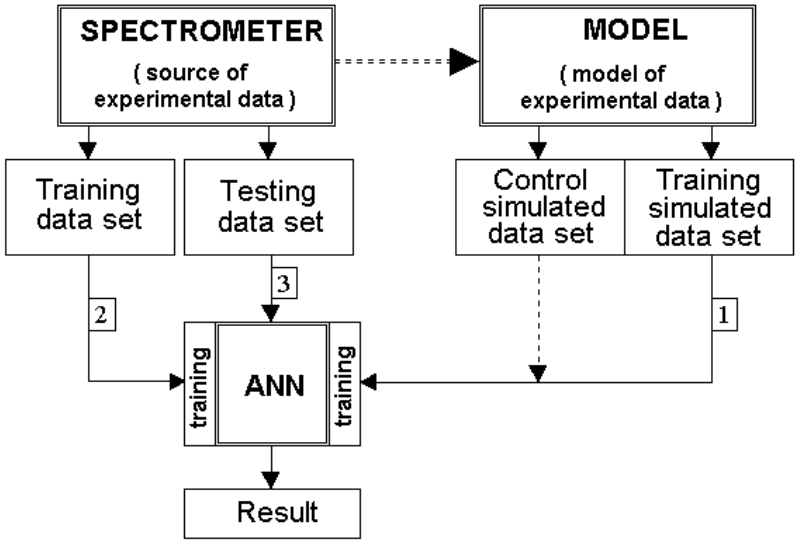


Figure 5.7. The algorithm of ANN training in the case of lack of experimental data.

5.5.3. Selection of optimal neuron number in hidden layers

The problem of optimal neuron number selection is still unsolved in the theory of multilayer networks. That is why we use the following empirical scheme to obtain an estimation of optimal number of neurons. Let us denote the target function F as a mean error of operation of ANN on a test set. This F can be considered as a function of neuron number in hidden layers. Now the problem of finding optimal neuron number turns to the problem of two-parametrical minimization of function F. In the current work this minimization was carried out by two standard procedures: scanning of parameter space to obtaining initial estimation and then sharpening of solution by the method of local variations. It should be mentioned that each value of F was calculated as a mean of 10 (for estimation) or 20 (for accurate solution) experiment.

The search of the optimal structure of a neural network was carried out. The network with 17 neurons in layer 1 and 4 neurons in layer 2 showed the best results.

The proposed method of neural network spectra analysis was tested on spectra of Cs 25μg/l water solutions. As a result of application of neural network RMSE of concentration

definition was reduced from 9.1% (regression approach) to 8.4 %. Moreover, the time of neural processing is much smaller, then by using of conventional methods of analysis. However, it should be mentioned, that the neural network should be retrained if a new element is to be analyzed, or if a new dye is used.

5.6. Estimation of the measurement errors

To accomplish the statistical analysis first of all it is necessary to find out a sort of distribution of measurement errors. After that, statistical characteristics of results are determined.

The distribution of recognized concentration (Fig. 5.8) was built using histograms and polygons of distributions for definition the sort of errors.

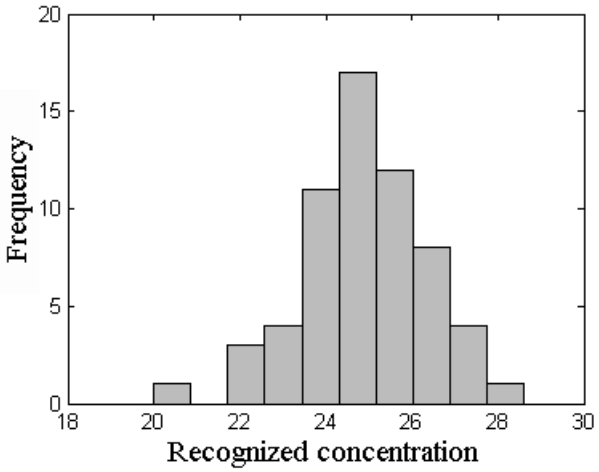


Figure 5.8. The histogram of recognized concentrations of Cs shows Gaussian behavior

The data from 61 spectra were obtained on Cs solutions with the constant concentration of 25 μg/l. Beforehand two suppositions were made. The first is that function of errors distribution is smooth and that measured value itself is continuous. Moreover the claim of symmetry of the errors distribution could be made. This is concerned with the relative smallness of errors. Even if the transfer function of our measuring equipment is essentially non-linear, the changing of its steepness on the small length, corresponding to the errors value, cannot lead to noticeable skewness of the distribution. The following estimations of the measurement errors had been received (see Table 5.1).

On the assumption of available estimations of k and κ, with the help of topographic classification of mathematical models of distribution, it is possible to assert that our distribution belongs either to exponential distributions or to triangle ones.

Table 5.1. Statistical parameters of measurements distributions for ANN method.

Parameters		Concentration
Distribution center (mean)	$X_C \pm \Delta_{0.9}$ (X_C)	24.9 ± 0.53
Standard deviation	$\sigma \pm \Delta_{0.9}$ (σ)	2.05 ± 0.27
Kurtosis	$\varepsilon \pm \Delta_{0.9}$ (ε)	2.76 ± 0.75
Antikurtosis	$\kappa \pm \Delta_{0.9}$ (κ)	0.60 ± 0.09
Entropic coefficient	$k \pm \Delta_{0.9}$ (k)	2.02 ± 0.10

5.7. Discussion

From the review presented in the second part of this work it may be concluded that for last few years artificial neural networks were applied as analytical tool in chemistry and close areas. The motivation for it was the intention to achieve better features of the available analytical system. There are examples of successful application of this flexible and powerful tool in atomic (Apanasovich et al., 2001) and molecular (Gedova et al., 2002) spectroscopy as well as in electron microscopy (Wienke et al., 1995) and mass-spectrometry (Eghbaldar et al., 1998).

This study as far as we know is the first attempt to use ANN for analysis of absorption spectra obtained in a high resolution spectrometry. As a rule a very limited volume of experimental data is the main restriction for ANNs implementation in high resolution spectroscopy.

Due to a novel approach the feed-forward ANN was successfully applied for absorption spectra processing and concentration extraction. The multilayer perceptron with 17 neurons in the first layer and 4 neurons in the second layer proved to be a quite suitable tool for this application. The simulated spectra were used for initial training of the ANN. Finally the neural network was trained with real experimental data. In the frame of the discussed approach the main problem in practical ANN application was avoided.

The approach discussed above is not limited by the area of high resolution spectroscopy and may be applied for different architectures of ANNs. We suppose that neural networks based on RBF (Apanasovich et al., 2001) may be widely used for the solving of similar problems along with multilayer perceptrons.

5.8. Conclusions

The problems of data processing in high resolution laser spectroscopy were discussed. The structure of intracavity laser spectrometer was described. The absorption line of explored atoms was selected by tuning of the flash lamp pumped dye laser.

The application of ANN results in decreasing of measurements error. The obtained relative standard deviation for definition of small concentrations of cesium in water solutions is decreased with respect to standard processing methods and equals to 8.4 %. The processing using ANN is robust and more accurate then the conventional methods. Furthermore it is less time consuming in comparison with regressive analysis and other conventional procedures.

Acknowledgments. The experiments in the frame of this work were performed in the Laboratory of Laser Diagnostics of Plasma, Institute of Molecular and Atomic Physics, National Academy of Sciences, Belarus. The financial support of the Belarusian State University project 540/18 is also gratefully acknowledged.

REFERENCES

Almeida, F.C.; Opella, S.J. (**1997**) fd coat protein structure in membrane environments: structural dynamics of the loop between the hydrophobic trans-membrane helix and the amphipathic in-plane helix. *J. Mol. Biol.*, 270(3), 481-495.

Andrews, L.; Demidov, A.A. (**1999**) *Resonance energy transfer*. Wiley, New York.

Apanasovich, V.V.; Novikov, E.G.; Yatskou, M.M. (**2000**) Analysis of the decay kinetics of fluorescence of complex molecular systems using the Monte Carlo method. *J. Appl. Spectrosc.*, 67(5), 842-851.

Apanasovich, V.V.; Balakhontsev, A.Y.; Lutkovski, V.M.; Misakov, P.Y.; Nazarov, P.V. (**2001**) Application of neural networks in analysis of atomic emission spectra. *International Conference on Neural Networks and Artificial Intelligence (ICNNAI)*, Minsk, Belarus, 177-180.

Apell, H.J.; Karlish, S.J. (**2001**) Functional properties of Na, K-ATPase, and their structural implications, as detected with biophysical techniques. *J. Membr. Biol.*, 180(1), 1-9.

Arora, A.; Tamm, L.K. (**2001**) Biophysical approaches to membrane protein structure determination. *Curr. Opin. Struct. Biol.*, 11(5), 540-547.

Bashtovyy, D.; Marsh, D.; Hemminga, M.A.; Pali, T. (**2001**) Constrained modeling of spin-labeled major coat protein mutants from M13 bacteriophage in a phospholipid bilayer. *Protein Sci.*, 10(5), 979-987.

Beechem, J.M.; Brand, L. (**1985**) Time-resolved fluorescence of proteins. *Annu. Rev. Biochem.*, 54, 43-71.

Beechem, J.M.; Brand, L. (**1986**) Global analysis of fluorescence decay: applications to some unusual experimental and theoretical studies. *Photochem. Photobiol.*, 44(3), 323-329.

Beechem, J.M.; Haas, E. (**1989**) Simultaneous determination of intramolecular distance distributions and conformational dynamics by global analysis of energy transfer measurements. *Biophys. J.*, 55(6), 1225-1236.

Berberan-Santos, M.N.; Valeur, B. (**1991**) Fluorescence depolarization by electronic energy transfer in donor-acceptor pairs of like and unlike chromophores. *J. Chem. Phys.*, 95(11), 8048-8055.

Berney, C.; Danuser, G. (**2003**) FRET or no FRET: a quantitative comparison. *Biophys. J.*, 84(6), 3992-4010.

Bishop, C.M. (**1995**) *Neural Networks for Pattern Recognition*. Oxford University Press.

Blumen, A.; Klafter, J.; Zumofen, G. (**1986**) Influence of restricted geometries on the direct energy transfer. *J. Chem. Phys.*, 84(3), 1397-1401.

Bogusky, M.J.; Schiksnis, R.A.; Leo, G.C.; Opella, S.J. (**1987**) Protein backbone dynamics by solid-state and solution ^{15}N NMR spectroscopy. *J. Magn. Reson.*, 72(1), 186-190.

Bogusky, M.J.; Leo, G.C.; Opella, S.J. (**1988**) Comparison of the dynamics of the membrane-bound form of fd coat protein in micelles and in bilayers by solution and solid-state nitrogen-15 nuclear magnetic resonance spectroscopy. *Proteins*, 4(2), 123-130.

Burakov, V.S.; Apanasovich, V.V.; Isaevich, A.V.; Lutkovski, V.M.; Misakov, P.Y.; Nazarov, P.V. (**2002**) Data processing and estimation of measurements errors in intracavity laser spectroscopy. *Proc. SPIE*, 4749, 172-177.

Byrne, B.; Iwata, S. (**2002**) Membrane protein complexes. *Curr. Opin. Struct. Biol.*, 12(2), 239-243.

Chiras, D.D. (**2002**) *Human Biology*. 4th ed. Jones & Bartlett Pub.

Cybenko, G. (**1989**) Approximations by superpositions of sigmoidal functions. *Math. Contr. Signals Syst.*, 2, 303-314.

Dale, R.E.; Eisinger, J.; Blumberg, W.E. (**1979**) The orientational freedom of molecular probes. The orientation factor in intramolecular energy transfer. *Biophys. J.*, 26, 161-193.

Davenport, L.; Dale, R.E.; Bisby, R.H.; Cundall, R.B. (**1985**) Transverse location of the fluorescent probe 1,6-diphenyl-1,3,5-hexatriene in model lipid bilayer membrane systems by resonance excitation energy transfer. *Biochemistry*, 24(15), 4097-4108.

Demidov, A.A.; Borisov, A.Y. (**1993**) Computer simulation of energy migration in the C-phycocyanin of the blue-green algae Agmenellum Quadruplicatum. *Biophys. J.*, 64(5), 1375-1384.

Devlin, T.M., editor (**2006**) *Textbook of biochemistry: with clinical correlations*. 6th ed. Wiley-Liss, Hoboken, NJ.

Dewey, T.G.; Hammes, G.G. (**1980**) Calculation on fluorescence resonance energy transfer on surfaces. *Biophys. J.*, 32(3), 1023-1035.

dos Remedios, C.G.; Moens, P.D. (**1995**) Fluorescence resonance energy transfer spectroscopy is a reliable "ruler" for measuring structural changes in proteins. Dispelling the problem of the unknown orientation factor. *J. Struct. Biol.*, 115(2), 175-185.

Eghbaldar, A.; Forrest, T.P.; Cabrol-Bass, D. (**1998**) Development of neural networks for identification of structural features from mass spectral data *Analyt. Chim. Acta*, 359(3), 283-301.

Fernandes, F.; Loura, L.M.; Prieto, M.; Koehorst, R.B.M.; Spruijt, R.B.; Hemminga, M.A. (**2003**) Dependence of M13 major coat protein oligomerization and lateral segregation on bilayer composition. *Biophys. J.*, 85(4), 2430-2441.

Fernandes, F.; Loura, L.M.; Koehorst, R.B.M.; Spruijt, R.B.; Hemminga, M.A.; Fedorov, A.; Prieto, M. (**2004**) Quantification of protein-lipid selectivity using FRET: application to the M13 major coat protein. *Biophys. J.*, 87(1), 344-352.

Fisher, C.A.; Ryan, R.O. (**1999**) Lipid binding-induced conformational changes in the N-terminal domain of human apolipoprotein E. *J. Lipid Res.*, 40(1), 93-99.

Fleming, P.J.; Koppel, D.E.; Lau, A.L.; Strittmatter, P. (**1979**) Intramembrane position of the fluorescent tryptophanyl residue in membrane-bound cytochrome b5. *Biochemistry*, 18(24), 5458-5464.

Förster, T. (**1948**) Intermolecular energy migration and fluorescence. *Ann. Phys.*, 2, 55-75.

Förster, T. (**1965**) Delocalized excitation and energy transfer. *In Modern Quantum Chemistry*. Sinanoglu, O., editor. Academic Press, New York.

Fox, G.C.; Williams, R.D.; Messina, P.C. (**1994**) *Parallel Computing Works!* Morgan Kaufmann.

Frederix, P.; de Beer, E.L.; Hamelink, W.; Gerritsen, H.C. (**2002**) Dynamic Monte Carlo simulations to model FRET and photobleaching in systems with multiple donor-acceptor interactions. *J. Phys. Chem. B*, 106(26), 6793-6801.

Gedova, I.V.; Churina, I.V.; Dolenko, S.A.; Dolenko, T.A.; Fadeev, V.V.; Persiantsev, I.G. (**2002**) New opportunities in solution of inverse problems in laser spectroscopy due to application of artificial neural networks. *Proc. SPIE*, 4749, 157-166.

Gennis, R.B. (**1989**) *Biomembranes: molecular structure and functon*. Cantor, C. R., editor. Springer-Verlag, New York.

Glaubitz, C.; Grobner, G.; Watts, A. (**2000**) Structural and orientational information of the membrane embedded M13 coat protein by (13)C-MAS NMR spectroscopy. *Biochim. Biophys. Acta*, 1463(1), 151-161.

Gustiananda, M.; Liggins, J.R.; Cummins, P.L.; Gready, J.E. (**2004**) Conformation of prion protein repeat peptides probed by FRET measurements and molecular dynamics simulations. *Biophys. J.*, 86(4), 2467-2483.

Hagan, M.T.; Menhaj, M. (**1994**) Training feedforward networks with the marquardt algorithm. *IEEE Trans. Neural Networks*, 5(6), 989-993.

Hemminga, M.A.; Sanders, J.C.; Wolfs, C.J.A.M.; Spruijt, R.B. (**1993**) Lipid-protein interactions involved in bacteriophage M13 infection. *In Protein-Lipid Interactions: new comprehensive biochemistry*. Watts, A., editor. Elsevier, Amsterdam. 191-212.

Henry, G.D.; Weiner, J.H.; Sykes, B.D. (**1987**) Backbone dynamics of a model membrane protein: assignment of the carbonyl carbon ^{13}C NMR resonances in detergent-solubilized M13 coat protein. *Biochemistry*, 26(12), 3619-3626.

Henry, G.D.; Sykes, B.D. (**1992**) Assignment of amide ^{1}H and ^{15}N NMR resonances in detergent-solubilized M13 coat protein: a model for the coat protein dimer. *Biochemistry*, 31(23), 5284-5297.

Hesselink, R.W.; Koehorst, R.B.M.; Nazarov, P.V.; Hemminga, M.A. (**2005**) Membrane-bound peptides mimicking transmembrane Vph1p helix 7 of yeast V-ATPase: a spectroscopic and polarity mismatch study. *Biochim. Biophys. Acta*, 1716(2), 137-145.

Hiriayama, S.; Sakai, Y.; Ghiggino, K.P.; Smith, T.A. (**1990**) The application of a simple deconvolution method to the analysis of stretched exponential fluorescence decay functions. *J. Photochem. Photobiol. A*, 52, 27-38.

Hornik, K.; Stinchcombe, M.; White, H. (**1989**) Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366.

Kale, U.; Voigtman, E. (**1995**) Signal processing of transient atomic absorption signals. *Spectrochim. Acta B*, 50(12), 1531-1541.

Kamal, J.K.; Behere, D.V. (**2002**) Spectroscopic studies on human serum albumin and methemalbumin: optical, steady-state, and picosecond time-resolved fluorescence studies, and kinetics of substrate oxidation by methemalbumin. *J. Biol. Inorg. Chem.*, 7(3), 273-283.

Karmazyn, M.; Sawyer, M.; Fliegel, L. (**2005**) The Na(+)/H(+) exchanger: a target for cardiac therapeutic intervention. *Curr. Drug Targets Cardiovasc. Haematol. Disord.*, 5(4), 323-335.

Keller, R.A.; Ambrose, W.P.; Goodwin, P.M.; Jett, J.H.; Martin, J.C.; Wu, M. (**1996**) Single-molecule fluorescence analysis in solution *Appl. Spectrsc.*, 50(7), 12A-32A.

Killian, J.A. (**2003**) Synthetic peptides as models for intrinsic membrane proteins. *FEBS Lett.*, 555(1), 134-138.

Knox, R.S.; van Amerongen, H. (**2002**) Refractive index dependence of the Förster resonance excitation transfer rate. *J. Phys. Chem. B*, 106, 5289-5293.

Koehorst, R.B.M.; Spruijt, R.B.; Vergeldt, F.J.; Hemminga, M.A. (**2004**) Lipid bilayer topology of the transmembrane α-helix of M13 Major coat protein and bilayer polarity profile by site-directed fluorescence spectroscopy. *Biophys. J.*, 87(3), 1445-1455.

Kohonen, T. (**1984**) *Self–organization and associative memory*. Springer-Verlag, Berlin.

Kolen, J.F.; Kremer, S.C. (**2001**) *A field guide to dynamical recurrent networks*. Wiley-IEEE Press.

Kolmogorov, A.N. (**1946**) On substantiation of the method of least squares. *Uspekhi Matematicheskikh Nauk (in Russian)*, 1(1), 57–70.

Lakey, J.H.; Duche, D.; Gonzalez-Manas, J.M.; Baty, D.; Pattus, F. (**1993**) Fluorescence energy transfer distance measurements. The hydrophobic helical hairpin of colicin A in the membrane bound state. *J. Mol. Biol.*, 230(3), 1055-1067.

Lakowicz, J.R. (**1999**) *Principles of fluorescence spectroscopy*. Kluwer Academic/Plenum Publishers, New York.

Lakshmikanth, G.S.; Sridevi, K.; Krishnamoorthy, G.; Udgaonkar, J.B. (**2001**) Structure is lost incrementally during the unfolding of barstar. *Nat. Struct. Biol.*, 8(9), 799-804.

Leo, G.C.; Colnago, L.A.; Valentine, K.G.; Opella, S.J. (**1987**) Dynamics of fd coat protein in lipid bilayers. *Biochemistry*, 26(3), 854-862.

Li, M.; Reddy, L.G.; Bennett, R.; Silva, N.D., Jr.; Jones, L.R.; Thomas, D.D. (**1999**) A fluorescence energy transfer method for analyzing protein oligomeric structure: application to phospholamban. *Biophys. J.*, 76(5), 2587-2599.

Loura, L.M.; Fedorov, A.; Prieto, M. (**1996**) Resonance energy transfer in a model system of membranes: application to gel and liquid crystalline phases. *Biophys. J.*, 71(4), 1823-1836.

Loura, L.M.; Fedorov, A.; Prieto, M. (**2001**) Fluid-fluid membrane microheterogeneity: a fluorescence resonance energy transfer study. *Biophys. J.*, 80(2), 776-788.

Low, A.M.; Kelton, W.D. (**2000**) *Simulation modeling and analysis*. 3 ed. McGraw-Hill.

Marassi, F.M.; Opella, S.J. (**2003**) Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints. *Protein Sci.*, 12(3), 403-411.

Marvin, D.A.; Hohn, B. (**1969**) Filamentous bacterial viruses. *Bacteriol Rev*, 33(2), 172-209.

Marvin, D.A.; Hale, R.D.; Nave, C.; Helmer-Citterich, M. (**1994**) Molecular models and structural comparisons of native and mutant class I filamentous bacteriophages Ff (fd, f1, M13), If1 and IKe. *J. Mol. Biol.*, 235(1), 260-286.

Marvin, D.A. (**1998**) Filamentous phage structure, infection and assembly. *Curr. Opin. Struct. Biol.*, 8(2), 150-158.

McDonnell, P.A.; Shon, K.; Kim, Y.; Opella, S.J. (**1993**) fd coat protein structure in membrane environments. *J. Mol. Biol.*, 233(3), 447-463.

Meijer, A.B.; Spruijt, R.B.; Wolfs, C.J.; Hemminga, M.A. (**2001a**) Membrane-anchoring interactions of M13 major coat protein. *Biochemistry*, 40(30), 8815-8820.

Meijer, A.B.; Spruijt, R.B.; Wolfs, C.J.; Hemminga, M.A. (**2001b**) Configurations of the N-terminal amphipathic domain of the membrane-bound M13 major coat protein. *Biochemistry*, 40(16), 5081-5086.

Mendes, P.; Kell, D. (**1998**) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10), 869-883.

Nazarov, P.V.; Popleteev, A.M.; Lutkovski, V.M. (**2002**) Identification of processes and systems using parallel simulation modeling and neural network approximation. *International Conference "Information systems and technologies"*, Minsk, Belarus, 142-146.

Nazarov, P.V.; Apanasovich, V.V.; Lutkovski, V.M.; Yatskou, M.M.; Koehorst, R.B.M.; Hemminga, M.A. (**2004**) Artificial neural network modification of simulation-based fitting: application to a protein-lipid system. *J. Chem. Inf. Comput. Sci.*, 44(2), 568-574.

Nazarov, P.V.; Koehorst, R.B.M.; Vos, W.L.; Apanasovich, V.V.; Hemminga, M.A. (**2006**) FRET study of membrane proteins: simulation-based fitting for analysis of membrane protein embedment and association. *Biophys. J.*, 91(2), 454-466.

Nelder, J.A.; Mead, R. (**1965**) A simplex method for function minimization. *Computer J.*, 7(4), 308-313.

Papavoine, C.H.; Remerowski, M.L.; Horstink, L.M.; Konings, R.N.; Hilbers, C.W.; van de Ven, F.J. (**1997**) Backbone dynamics of the major coat protein of bacteriophage M13 in detergent micelles by 15N nuclear magnetic resonance relaxation measurements using the model-free approach and reduced spectral density mapping. *Biochemistry*, 36(13), 4015-4026.

Papavoine, C.H.; Christiaans, B.E.; Folmer, R.H.; Konings, R.N.; Hilbers, C.W. (**1998**) Solution structure of the M13 major coat protein in detergent micelles: a basis for a model of phage assembly involving specific residues. *J. Mol. Biol.*, 282(2), 401-419.

Pollard, T.D.; Borisy, G.G. (**2003**) Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112(4), 453-465.

Pulido, A.; Ruisanchez, I.; Rius, F.X. (**1999**) Radial basis functions applied to the classification of UV-visible spectra. *Analyt. Chim. Acta*, 388(3), 273-281.

Ramadan, Z.; Song, X.-H.; Hopke, P.K.; Johnson, M.J.; Scow, K.M. (**2001**) Variable selection in classification of environmental soil samples for partial least square and neural network models. *Analyt. Chim. Acta*, 446(1), 231-242.

Rao, M.; Mayor, S. (**2005**) Use of Forster's resonance energy transfer microscopy to study lipid rafts. *Biochim. Biophys. Acta*, 1746(3), 221-233.

Ren, B.; Gao, F.; Tong, Z.; Yan, Y. (**1999**) Solvent polarity scale on the fluorescence spectra of a dansyl monomer copolymerizable in aqueous media. *Chem. Phys. Lett.*, 307(1), 55-61.

Ruisanchez, I.; Lozano, J.; Larrechi, M.S.; Rius, F.X.; Zupan, J. (**1997**) On-line automated analytical signal diagnosis in sequential injection analysis systems using artificial neural networks. *Analyt. Chim. Acta*, 348(1-3),

Safavi, A.; Absalan, G.; Maesum, S. (**2001**) Simultaneous determination of V(IV) and Fe(II) as catalyst using "neural networks" through a single catalytic kinetic run. *Analyt. Chim. Acta*, 432(2), 229-233.

Sambrook, J.; Fritsch, E.F.; Maniatis, T. (**1989**) *Molecular cloning*. 2nd ed. Cold Spring Harbor Laboratory Press, New York.

Schulz, H.; Derrick, M.; Stulik, D. (**1995**) Simple encoding of infrared spectra for pattern recognition Part 2. Neural network approach using back-propagation and associative Hopfield memory. *Analyt. Chim. Acta*, 316(2), 145-159.

Sparr, E.; Ash, W.L.; Nazarov, P.V.; Rijkers, D.T.; Hemminga, M.A.; Tieleman, D.P.; Killian, J.A. (**2005**) Self-association of transmembrane α-helices in model membranes: importance of helix orientation and role of hydrophobic mismatch. *J. Biol. Chem.*, 280(47), 39324-39331.

Spruijt, R.B.; Wolfs, C.J.; Hemminga, M.A. (**1989**) Aggregation-related conformational change of the membrane-associated coat protein of bacteriophage M13. *Biochemistry*, 28(23), 9158-9165.

Spruijt, R.B.; Wolfs, C.J.; Verver, J.W.; Hemminga, M.A. (**1996**) Accessibility and environment probing using cysteine residues introduced along the putative transmembrane domain of the major coat protein of bacteriophage M13. *Biochemistry*, 35(32), 10383-10391.

Spruijt, R.B.; Meijer, A.B.; Wolfs, C.J.; Hemminga, M.A. (**2000**) Localization and rearrangement modulation of the N-terminal arm of the membrane-bound major coat protein of bacteriophage M13. *Biochim. Biophys. Acta*, 1509(1-2), 311-323.

Spruijt, R.B.; Wolfs, C.J.; Hemminga, M.A. (**2004**) Membrane assembly of M13 major coat protein: evidence for a structural adaptation in the hinge region and a tilted transmembrane domain. *Biochemistry*, 43(44), 13972-13980.

Stegemann, J.A.; Buenfeld, N.R. (**1999**) A glossary of basic neural network terminology for regression problems. *Neural Comput. Appl.*, 8(4), 290-296.

Stopar, D.; Spruijt, R.B.; Wolfs, C.J.; Hemminga, M.A. (**2002**) Structural characterization of bacteriophage M13 solubilization by amphiphiles. *Biochim. Biophys. Acta*, 1594(1), 54-63.

Stopar, D.; Spruijt, R.B.; Wolfs, C.J.; Hemminga, M.A. (**2003**) Protein-lipid interactions of bacteriophage M13 major coat protein. *Biochim. Biophys. Acta*, 1611(1-2), 5-15.

Stopar, D.; Strancar, J.; Spruijt, R.B.; Hemminga, M.A. (**2005**) Exploring the local conformational space of a membrane protein by site-directed spin labeling. *J. Chem. Inf. Model.*, 45(6), 1621-1627.

Stopar, D.; Spruijt, R.B.; Hemminga, M.A. (**2006a**) Anchoring mechanisms of membrane-associated M13 major coat protein. *Chem. Phys. Lipids*, 141, 83-93.

Stopar, D.; Strancar, J.; Spruijt, R.B.; Hemminga, M.A. (**2006b**) Motional restrictions of membrane proteins: A site-directed spin labeling study. *Biophys. J.*, in press.

Strancar, J.; Koklic, T.; Arsov, Z.; Filipic, B.; Stopar, D. (**2005**) Spin label EPR-based characterization of biosystem complexity. *J. Chem. Inf. Model.*, 45, 394-406.

Stryer, L.; Haugland, R.P. (**1967**) Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci. USA*, 58(2), 719-726.

Stryer, L. (**1978**) Fluorescence energy transfer as a spectroscopic ruler. *Annu. Rev. Biochem.*, 47, 819-846.

Tetko, I.V.; Livingstone, D.J.; Luik, A.I. (**1995**) Neural network studies 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.*, 35, 826-833.

Tomassini, M. (**1999**) Parallel and distributed evolutionary algorithms: a review. *In Evolutionary Algorithms in Engineering and Computer Science*. Miettinen, K., editor. John Wiley & Sons ltd, New York. 113-133.

Torres, J.; Stevens, T.J.; Samso, M. (**2003**) Membrane proteins: the 'Wild West' of structural biology. *Trends Biochem. Sci.*, 28(3), 137-144.

Valeur, B. (**2001**) *Molecular fluorescence: principles and applications*. Wiley-VCH.

van Amerongen, H.; Valkunas, L.; van Grondelle, R. (**2000**) *Photosynthetic Excitons* World Scientific Publishing Company.

Vanderkooi, J.M.; Ierokomas, A.; Nakamura, H.; Martonosi, A. (**1977**) Fluorescence energy transfer between Ca2+ transport ATPase molecules in artificial membranes. *Biochemistry*, 16(7), 1262-1267.

Vanderkooi, J.M. (**2002**) Tryptophan phosphorescence from proteins at room temperature. *In Topics in fluorescence spectroscopy. Biochemical applications*. Lakowicz, J. R., editor. Kluwer Academic Publishers, New York.

Vassar, R. (**2002**) Beta-secretase (BACE) as a drug target for Alzheimer's disease. *Adv. Drug Deliv. Rev.*, 54(12), 1589-1602.

Vogel, S.S.; Thaler, C.; Koushik, S.V. (**2006**) Fanciful FRET. *Sci. STKE*, 2006(331), re2.

Vos, W.L.; Koehorst, R.B.M.; Spruijt, R.B.; Hemminga, M.A. (**2005**) Membrane-bound conformation of M13 major coat protein: a structure validation through FRET-derived constraints. *J. Biol. Chem.*, 280(46), 38522-38527.

Walczak, B. (**1996**) Neural networks with robust backpropagation learning algorithm. *Analyt. Chim. Acta*, 322(1), 21-29.

Walczak, B.; Massart, D.L. (**1996a**) Application of radial basis functions – Partial Least Squares to non-linear pattern recognition problems: diagnostics of process faults. *Analyt. Chim. Acta*, 331(3), 187-193.

Walczak, B.; Massart, D.L. (**1996b**) The radial basis functions – partial least squares approach as a flexible non-linear regression technique. *Analyt. Chim. Acta*, 331(3), 177-185.

Wang, H.; Zhou, Y.; Li, Q.; Chen, X.; Hu, Z. (**2001**) Optimization of on-line microwave flow injection analysis system by artificial neural networks for the determination of ruthenium. *Analyt. Chim. Acta*, 429(2), 207-213.

Wasserman, P.D. (**1989**) *Neural computing: theory and practice*. Van Nostrand Reinhold, New York.

Weinglass, A.B.; Whitelegge, J.P.; Kaback, H.R. (**2004**) Integrating mass spectrometry into membrane protein drug discovery. *Curr. Opin. Drug Discov. Devel.*, 7(5), 589-599.

White, S.H.; Wimley, W.C. (**1999**) Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, 28, 319-365.

White, S.H. (**2004**) The progress of membrane protein structure determination. *Protein Sci.*, 13(7), 1948-1949.

Whiteside, P.J. (**1988**) *Atomic absorption data book*. Black Bear Press Limited, Cambridge.

Wienke, D.; Xie, Y.; Hopke, P.K. (**1995**) Classification of airborne particles by analytical scanning electron microscopy imaging and a modified Kohonen neural network (3MAP). *Analyt. Chim. Acta*, 310(1), 1-14.

Wolber, P.K.; Hudson, B.S. (**1979**) An analytic solution to the Forster energy transfer problem in two dimensions. *Biophys. J.*, 28(2), 197-210.

Wolkers, W.F.; Spruijt, R.B.; Kaan, A.; Konings, R.N.; Hemminga, M.A. (**1997**) Conventional and saturation-transfer EPR of spin-labeled mutant bacteriophage M13 coat protein in phospholipid bilayers. *Biochim. Biophys. Acta*, 1327(1), 5-16.

Yatskou, M.M.; Donker, H.; Koehorst, R.B.M.; van Hoek, A.; Schaafsma, T.J. (**2001a**) A study of energy transfer processes in zinc-porphyrin films using Monte Carlo simulation of fluorescence decay. *Chem. Phys. Lett.*, 345(1), 141-150.

Yatskou, M.M.; Donker, H.; Novikov, E.G.; Koehorst, R.B.M.; van Hoek, A.; Apanasovich, V.V.; Schaafsma, T.J. (**2001b**) Nonisotropic excitation energy transport in organized molecular systems: Monte Carlo simulation-based analysis of time-resolved fluorescence. *J. Phys. Chem. A*, 105(41), 9498-9508.

Yatskou, M.M.; Meyer, M.; Huber, S.; Pfenninger, M.; Calzaferri, G. (**2003**) Electronic excitation energy migration in a photonic dye-zeolite antenna. *Chem. Phys. Chem.*, 4, 567-587.

SUMMARY

Membrane proteins play an important role in almost all cell activities. However, the characterization of the structure of membrane proteins in lipid bilayers is still at the frontier of structural biology. While 30-40% of all proteins are situated at or in membranes, yet less than 1% of the known protein structures are of membrane proteins. The complexity and delicacy of membrane-protein systems impedes the application of standard methods of protein study, such as X-ray crystallography and NMR. Furthermore, these techniques are mainly aimed at short-range structural information, and seem to be not very useful for the study of long-range interactions, for instance in the case of protein assemblies. These factors urge to find other biophysical approaches to study proteins incorporated into lipid bilayers. A successful alternative is Förster (or fluorescence) resonance energy transfer (FRET) spectroscopy in combination with site-directed labeling with fluorescent probes. This technique provides distance information within a range of 10-100 Å, which is sufficient to study the structure of membrane proteins and their complexes. The crucial point in the extraction of structural information from FRET data is an advanced and robust data analysis. This work is devoted to the development of such methods for analysis of fluorescence data, based on simulation modeling, global analysis and artificial neural networks. Especially the advances and problems of the simulation-based fitting (SBF) approach to fluorescence data analysis are considered. The methodologies of global analysis and SBF are applied to obtain information about the position, aggregation and structure of M13 major coat protein in DOPC:DOPG vesicles. The resulting physical parameters, that describe the embedment and orientation of the protein in the membrane, such as protein-protein aggregation, protein depth, tilt angle, and tilt direction, are in good accordance with previously reported values. Based on the FRET data, it was found that M13 major coat protein (having 50 amino acid residues) in its bilayer conformation could be described as a single α-helix between amino acid positions 10-46. Additional work was performed on the methodological aspects of improving the SBF data analysis technique. Here it is proposed to use an artificial neural network to speed up the parameter identification and to make the process of fitting less sensitive to noise. The main idea of this method is the substitution of a time-consuming simulation model by an artificial neural network, specifically a multi-layer perceptron. The method results in a speeding up of the simulation by about a factor of 10^4 for the developed FRET model.

SAMENVATTING

Membraaneiwitten spelen een belangrijke rol is vrijwel alle activiteiten van een biologische cel. Echter, de bepaling van de structuur van membraaneiwitten in lipidebilagen is nog steeds een onontgonnen gebied van de structurele biologie. Terwijl 30-40% van alle eiwitten op of in het membraan zit, is slechts 1% van de bekende eiwitstructuren afkomstig van membraaneiwitten. De complexiteit en delicaatheid van de membraaneiwitsystemen belemmert de toepassing van standaardtechnieken voor eiwitonderzoek, zoals Röntgendiffractie en NMR. Bovendien zijn deze technieken gericht op de bepaling van structuurinformatie op korte afstanden, waardoor ze niet bijzonder geschikt zijn voor de bestudering van lange afstandsinteracties, bijvoorbeeld in het geval van eiwitaggregaten. Deze factoren nodigen uit om andere biofysische benaderingen te vinden voor de bestudering van eiwitten die in lipidebilagen zijn geïncorporeerd. Een succesvol alternatief is Förster (of: fluorescentie) resonantie-energieoverdracht ("fluorescence resonance energy transfer", FRET) spectroscopie in combinatie met plaatsgerichte labeling met fluorescente kleurstofmoleculen. Deze techniek is in staat om afstandsinformatie te geven op een schaal van 10 tot 100 Å, wat voldoende is voor de bestudering van de structuur van membraaneiwitten en hun complexen. Van doorslaggevend belang voor het verkrijgen van structurele informatie uit de FRET-meetgegevens is een geavanceerde en robuuste data-analyse. Het werk dat in dit proefschrift wordt beschreven, is gewijd aan de ontwikkeling van zulke methoden voor de analyse van fluorescentiemeetgegevens, gebaseerd op simulatiemodellering, globale analyse en kunstmatige neuronale netwerken. In het bijzonder wordt aandacht besteed aan de ontwikkeling en problemen van een methode die uitgaat van het zo goed mogelijk aanpassen van een simulatiemodel op fluorescentiemeetgegevens ("simulation-based fitting", SBF). De methodologie van de globale analyse en SBF zijn toegepast om informatie te verkrijgen over de positie, aggregatie en structuur van het manteleiwit van de bacteriofaag M13 in membraanblaasjes gemaakt van de fosfolipiden DOPC en DOPG. De gevonden fysische parameters die de inbedding en oriëntatie van het eiwit in het membraan beschrijven, zoals de eiwitaggregatie, eiwitdiepte, tilthoek en tiltrichting, zijn in goede overeenstemming met eerder gerapporteerde waarden. Uit de FRET-meetgegevens wordt geconcludeerd dat M13-manteleiwit (dat bestaat uit 50 aminozuurresiduen) in de bilaagconformatie kan worden beschreven als een enkelvoudige α-helix die loopt van residu 10 tot 46. Aanvullend methodologisch onderzoek is uitgevoerd om de SBF-methode verder te verbeteren. In dit verband wordt voorgesteld om een kunstmatige neuronaal netwerk toe te passen om de bepaling van de parameters te versnellen en om het proces van de aanpassing van de

meetgegevens minder gevoelig te maken voor ruis. De kern van deze methode is om het tijdrovende simulatiemodel te vervangen door een kunstmatige neuronaal netwerk, in het bijzonder een meerlaags perceptron. De methode resulteert in een versnelling van de simulatie met een factor 10^4 voor het ontwikkelde FRET-model.

АБАГУЛЬНЕННЕ

Мембранныя пратэіны выконваюць вельмі важную ролю практычна ва ўсіх клеткавых працэсах. Аднак, нягледзячы на важнасць даследвання мембранных пратэінаў, вызначэнне іх структуры з'яўляецца да гэтага часу вельмі складанай задачай. Складанасць і хрупкасць сістэмы пратэін-мембрана ў значнай ступені перашкоджваюць выкарыстоўванню такіх стандартных метадаў, як рэнтгенаўская крысталаграфія і ЯМР. Гэтыя фактары прымушаюць да пошукаў іншых падыходаў да вывучэння пратэінаў, што знаходзяцца ў ліпідным бісloі. Альтэрнатыўнай методыкай, якая паспяхова выкарыстоўваецца зараз, з'яўляецца флюарысцэнтная спектраскапія рэзананснага пераносу энэргіі (РПЭ) ў звязку з сайт-спецыфічным укараненнем метак. Гэты метад дазваляе атрымліваць інфармацыю аб адлегласцях у дыапазоне 10-100 Å, што з'яўляецца дастатковым для вывучэння як саміх мембранных пратэінаў, так і іх комплексаў. Вельмі важным пры атрыманні структурнай інфармацыі з РПЭ экперымента з'яўляецца выкарыстанне эфектыўнага аналізу дадзеных. Мэта гэтай работы – выпрацоўка і апрабацыя метада аналізу дадзенных флюарысцэнтнай спектраскапіі РПЭ пры выкарысанні апарата імітацыйнага мадэлявання, глабальнага аналізу і штучных няйронных сетак. Асаблівая ўвага ў рабоце надаецца праблеме фітынгу дадзеных мадэллю – дадатковыя даследванні былі праведзены у напрамку развіцця мэтадалагічнага аспэкту гэтага падыхода. У рабоце прапанавана выкарыстоўваць штучную няйронную сетку для паскарэння ідэнтыфікацыі параметраў пры фітынге і паніжэння ўплыву стахастычных фактараў. Асноўная ідэя метада – замена рэсурсаёмістай імітацыоннай мадэлі на спецыяльна навучанную няйронную сетку. У якасці няйроннай сеткі было прапанавана скарыстаць шматслаёвы персэптрон, які з'яўляецца ўніверсальным апраксіматарам. У выпадку разглядаемай у рабоце мадэлі пераноса энэргіі, паскарэнне мадэлявання пры замене складае прыкладна 10^4 разоў. Метадалогія глабальнага аналізу і фітынг мадэллю былі выкарыстаны для атрымання інфармацыі пра павозdіны базавага абалонкавага пратэіну бактэрыафага M13 у ліпідных везікулах. Знойдзеныя параметры, такія як: агрэгаванасць, глыбіня, вугал і напрамак нахілу, што апісваюць паводзіны пратэіна ў мембране, добра стасуюцца з раней апублікаванымі дадзенымі. Глабальны аналіз дадзеных РПЭ эксперыментаў выявіў, што пратэін бактэрыафага M13 (які мае 50 амінакіслотных рэштак) у сваёй мембраннай канфармацыі можа быць апісаны адной α-скруткай паміж пазіцыямі 10-46.

РЕЗЮМЕ

Мембранные протеины играют исключительно важную роль в подавляющем большинстве клеточных процессах. Однако, несмотря на чрезвычайную значимость исследования мембранных протеинов, определение их структуры является очень сложной и до сих пор далекой от своего разрешения задачей. Сложность и хрупкость системы протеин-мембрана в значительной степени затрудняют применение таких стандартных методов определения структуры, как рентгеновская кристаллография и ЯМР. В этой ситуации необходимо искать другие подходы к изучению протеинов встроенных в липидный бислой. Успешно применяемой альтернативой является флуоресцентная спектроскопия резонансного переноса энергии (РПЭ) в комбинации с сайт-специфическим внедрением меток. Ключевым моментом при получении структурной информации из РПЭ эксперимента является применение эффективных процедур анализа данных. Целью данной работы является разработка и апробация метода анализа данных флуоресцентной спектроскопии РПЭ, основанного на использовании аппарата имитационного моделирования, глобального анализа и искусственных нейронных сетей. Особое внимание в работе уделено проблеме анализа данных с помощью имитационного моделирования, в частности были проведены исследования с целью развития методологического аспекта приближения с использованием имитационной модели. В работе предложено применять искусственную нейронную сеть для ускорения идентификации параметров и снижения влияние стохастических факторов. Основная идея метода заключается в замене ресурсоемкой имитационной модели специально обученной нейронной сетью. В качестве нейронной сети предложено использовать многослойный персептрон, являющийся универсальным аппроксиматором. В случае рассматриваемой в работе модели переноса энергии, ускорение при такой замене составило порядка 10^4 раз. Методология глобального анализа и имитационного моделирования были применены для получения информации о поведении основного оболочечного протеина бактериофага M13 в липидных везикулах. Найдены параметры, описывающие его положение в мембране: аггрегированость, глубина, угол и направление наклона. Полученные результаты находятся в хорошем соответствии с ранее опубликованными данными. На основании глобального анализа данных РПЭ экспериментов было выяснено, что оболочечный протеин бактериофага M13 (имеет 50 аминокислотных остатков) в своей мембранной конформации может быть описан одной α-спиралью между позициями 10-46.

ACKNOWLEDGMENTS

This dissertation represents the result of several years of work. Obviously, this result would never have reached without the assistance of other people. Due to the special set-up of my sandwich PhD project the number of those, who have interacted with me and shared their experience, doubles. Therefore I would like to thank all with whom I had scientific-related discussions during these years. Especially I should mention those who are listed below.

First, I would like to express my eternal gratitude to my supervisor – Marcus Hemminga (Laboratory of Biophysics, Wageningen University), whose outstanding motivation, pedagogic skill and constant attention allowed me to start and complete this work. No one, except for my parents, has spent such an enormous amount of time on me.

Many thanks to my closest colleague in Wageningen – Rob Koehorst, who provided me with a lot of knowledge about fluorescence and photo-physical processes. I will always remember our discussions from which I got more on these subjects than from books.

I owe my gratitude to the secretary of the Laboratory of Biophysics – Netty Hoefakker, who arranged my visits to Wageningen and helped a lot in the organization of my doctoral defense.

I want to mention the Head of the Laboratory of Biophysics, Prof. Herbert van Amerongen. Herbert, being the promoter of the PhD project and an advanced expert in photophysics, made a significant and indispensable input in editing this thesis.

Thanks to all my colleagues in the Laboratory of Biophysics, Wageningen University, for their kindness, openness and readiness to help. Werner Vos, Afonso Duarte, Arie van Hoek, Ruud Spruijt, Bart van Oort, Cor Wolfs, Cor Dijkema, Edo Gerkema, Carel Windt, Frank Vergeldt and Natalia Homan – special thanks to you.

During the periods of my work at the Department of Systems Analysis, Belarusian State University, I had the pleasure to work under the supervision of Prof. Vladimir Apanasovich, whose organizations skills allowed me to start the project. His deep knowledge of simulation of stochastic processes helped a lot during the work.

Special thanks to Vladimir Lutkovski (Belarusian State University), my first supervisor during my undergraduate and master study, who drew me into the area of data analysis and neural networks. He was the first person who trained me how “to be a scientist”.

I want to thank the older colleague and my predecessor at the PhD studentship in Wageningen, who has now become my good friend – Mikalai Yatskou, for his irreplaceable advices and for our fruitful discussions concerning simulation and analysis of energy transfer processes.

Warmest thanks and my love to my family – mother Tatiana, wife Irina and son Andrei for their support, love and understanding.

I also want to thank my good friends and younger colleagues Marina Repich and Andrei Popleteev (Belarusian State University, now – Trento University, Italy) for their friendship and fruitful collaboration in the areas of computer simulation and programming.

Many thanks to Viktor Khutko (Suss MicroTech Systems, Germany) and Alexander Ivaniukovich (Trento University, Italy), my friends, who helped me a lot by sharing computational power of their computers and local networks during the study of simulation-based fitting methodology and data analysis.

Finally, thanks to my friend Aliaksei Pukin (Laboratory of Organic Chemistry, Wageningen University), who, being an advanced expert in chemistry and many languages (Russian, Belarusian, English, French, Spanish, Dutch, German, Italian, etc), always helped me when I faced a chemical or linguistic problem.

CURRICULUM VITAE

Petr Nazarov was born on the 1st of January 1978 in Minsk, Belarus, in the family of Vladimir Nazarov and Tatiana Beznos. His elementary studies were completed at the Lyceum of Belarusian State University (Minsk) in 1995. In 2000 he graduated from Radio Physics faculty of the Belarusian State University, and in 2001 he received his Master of Science degree. Starting from 1998 he studied and worked as laboratory assistant, engineer and junior research assistant at the Department of Systems Analysis (Radio Physics faculty, Belarusian State University) on the research projects concerned with data analysis, neural networks and simulation modeling. From February 2003 to December 2006 he carried out his PhD study in the Laboratory of Biophysics, Wageningen University, The Netherlands and the Department of Systems Analysis, Belarusian State University, Minsk, Belarus. He published 10 journal articles and has 15 publications in proceedings of International Conferences.

PUBLICATIONS

Burakov, V.S.; Apanasovich, V.V.; Isaevich, A.V.; Lutkovski, V.M.; Misakov, P.Y.; Nazarov, P.V. (**2002**) Data processing and estimation of measurements errors in intracavity laser spectroscopy. *Proc. SPIE*, 4749, 172-177.

Hesselink, R.W.; Koehorst, R.B.M.; Nazarov, P.V.; Hemminga, M.A. (**2005**) Membrane-bound peptides mimicking transmembrane Vph1p helix 7 of yeast V-ATPase: a spectroscopic and polarity mismatch study. *Biochim. Biophys. Acta*, 1716(2), 137-145.

Isaevich, A.V.; Kozlovski, A.S.; Lutkovski, V.M.; Misakov, P.Y.; Nazarov, P.V. (**2001**) Increase of accuracy in the measurement of the elements in the atom emission spectral analysis. *Vesti NAN Belarusi (in Russian)*, (2), 80-85.

Isaevich, A.V.; Lutkovski, V.M.; Misakov, P.Y.; Nazarov, P.V. (**2003**) Determination of small amounts of elements by the atomic emission method with preliminary concentration of a sample. *J. Appl. Specrosc.*, 70(5), 672-676.

Nazarov, P.V.; Apanasovich, V.V.; Lutkovskaya, K.U.; Lutkovski, V.M.; Misakov, P.Y. (**2003a**) Neural network data analysis for intracavity laser spectroscopy. *Proc. SPIE*, 5135, 61-69.

Nazarov, P.V.; Apanasovich, V.V.; Lutkovski, V.M.; Hemminga, M.A.; Koehorst, R.B.M. (**2003b**) Neural network simulation of energy transfer processes in a membrane protein system. *Advances in Soft Computing: Neural Networks and Soft Computing*. Physica-Verlag, 873-878.

Nazarov, P.V.; Apanasovich, V.V.; Lutkovski, V.M.; Yatskou, M.M.; Koehorst, R.B.M.; Hemminga, M.A. (**2004**) Artificial neural network modification of simulation-based fitting: application to a protein-lipid system. *J. Chem. Inf. Comput. Sci.*, 44(2), 568-574.

Nazarov, P.V.; Koehorst, R.B.M.; Vos, W.L.; Apanasovich, V.V.; Hemminga, M.A. (**2006**) FRET study of membrane proteins: simulation-based fitting for analysis of membrane protein embedment and association. *Biophys. J.*, 91(2), 454-466.

Nazarov, P.V.; Koehorst, R.B.M.; Vos, W.L.; Apanasovich, V.V.; Hemminga, M.A. (**2007**) FRET study of membrane proteins: determination of the tilt and orientation of the N-terminal domain of M13 major coat protein. *Biophys. J.*, 92(4), in press.

Sparr, E.; Ash, W.L.; Nazarov, P.V.; Rijkers, D.T.; Hemminga, M.A.; Tieleman, D.P.; Killian, J.A. (**2005**) Self-association of transmembrane α-helices in model membranes: importance of helix orientation and role of hydrophobic mismatch. *J. Biol. Chem.*, 280(47), 39324-39331.