# Analysis of the Bacteriorhodopsin Photocycle by Singular Value Decomposition with Self-Modeling: A Critical Evaluation Using Realistic Simulated Data

**László Zimányi***

*Institute of Biophysics, Biological Research Center of the Hungarian Academy of Sciences, P.O.Box 521, Szeged, Hungary H-6701*

*Received: August 20, 2003; In Final Form: November 3, 2003*

Data matrices consisting of sample optical absorption as a function of wavelength and another variable, such as time, are decomposable using known matrix algebraic methods. The natural decomposition, based on the Beer−Lambert law, into the product of component absorbance and (time dependent) concentration matrices is usually not straightforward. Singular value decomposition yields orthonormal spectral and kinetic eigenvectors, with mathematical but not physical meaning. The connection of the two decompositions is explored with reference to the problem of the bacteriorhodopsin photocycle. The limitations and applicability of singular value decomposition with self-modeling is evaluated with known stoichiometric constraints on the intermediate kinetics and compared to other techniques applied to the same problem. The improved method of exponential fit assisted self-modeling is introduced and demonstrated on realistic simulated data.

## Introduction

Absorption spectroscopy has been widely used to investigate the function of various chromophoric systems under different experimental conditions, or as a function of time elapsed after an actinic light pulse. One of the examples is the photocycle of bacteriorhodopsin, a retinal protein in the cell membrane of *Halobacterium salinarum*. This molecule performs transmembrane active proton transfer following the light induced isomerization of the retinal, and the cyclic structural changes accompanying proton pumping are reflected in changes of the absorption spectrum, initially having a maximum at 570 nm in the light adapted state of the wild-type protein at room temperature (for recent reviews, see refs 2 and 3). Spectrally distinct intermediates follow each other roughly as J(625 nm), K(590 nm), L(545 nm), M(412 nm), N(563 nm), and O(630 nm), with the numbers indicating estimated maxima of the corresponding room-temperature absolute spectra.[4,5] The sequence of events and their interrelations comprising the photocycle have been the subject of debate over the last several decades. An optimal solution would consist of a photocycle scheme whose rate constants exhibit normal Arrhenius behavior, and which explains the observed dependence of the kinetics on such important factors as pH and actinic light intensity. The scheme should be complex enough to account for those molecular events taking place during the photocycle, whose existence is established by a large body of experimental data, including other spectroscopic techniques, most notably FTIR, structural investigations, and site directed mutagenesis.[6−11] In addition, the absorption spectra of the intermediates should possess expected properties based on experience with various visual and ion translocating retinal proteins: a broad, asymmetric, single main band and smaller peaks forming a plateau on the blue side of the main band.[12]

The dissection of the bacteriorhodopsin photocycle from kinetic absorption measurements is a difficult task for various reasons. The visible absorption bands of the intermediates strongly overlap, and their exact spectral shape is not known a priori. The determination of the photocycling ratio (PCR), i.e., the percentage of bacteriorhodopsin molecules entering the photocycle after a single laser flash may be ambiguous. The accumulation of the intermediates also overlap temporally, which results in an unknown mixture of intermediates present at any time before the completion of the photocycle. Intermediate substates with indistinguishable spectra may exist, and rapid equilibria between intermediate forms may develop so that the corresponding spectra cannot be separated. These complications may be present even if one assumes a homogeneous initial sample population. Further complications may arise if, as suggested by several investigators, the bacteriorhodopsin population is functionally heterogeneous or becomes heterogeneous during the photocycle, which would require parallel or branching photocycle schemes.[13−15]

Recently, Shrager and Hendler[16] reviewed the kinetic models for the bacteriorhodopsin photocycle and the various strategies applied to obtain such models from spectroscopic data. Strategies can be classified in several respects. The data can be directly fitted or first subjected to a matrix decomposition such as singular value decomposition (SVD) and the resulting component matrices analyzed. The fit can be multiexponential[17,18] or model-dependent.[19] A combination of these two procedures has helped to select potential models based on the shape of the exponential amplitude spectra in the analogous cases of rhodopsin and sensory rhodopsin.[20,21] A separate class of procedures attempts to determine the intermediate spectra first,[5,22−27] for instance by an adaptation of the chemometric method self-modeling (SM),[28] and obtain the kinetics as a straightforward step afterward. Photocycle schemes are then fitted to the kinetics, while the acceptable intermediate spectra are secured in the first step.[29] A global model fit, on the other hand, tries to determine the intermediate kinetics and spectra simultaneously in successive iterations. The mathematics of this latter method has been reviewed by several authors,[19,30] whereas the mathematics, applicability, and limitations of singular value decomposition

* To whom correspondence should be addressed. E-mail: zimanyi@nucleus.szbk.u-szeged.hu. Phone: +36 62 599 607. Fax: +36 62 433 133.

10.1021/jp0364809 CCC: $27.50 
Published on Web 03/04/2004

with self-modeling (SVD-SM) has not yet been thoroughly presented. In this article, I attempt to summarize the method of SVD-SM and its implementation to resolve kinetic absorption data characteristic of processes such as the bacteriorhodopsin photocycle. The analysis has implications regarding also the applicability of other methods, used by other investigations, to dissect the photocycle. A combination of a global multiexponential fit with self-modeling (exponential fit assisted self-modeling, EFASM) is shown to have additional advantages in correctly determining the stoichiometric sum of intermediate concentrations and to improve the search for the pure intermediate spectra.

## Theoretical

**Formulation of the Problem.** In absorption kinetics measurements at multiple wavelengths and discrete time delays after the actinic laser flash, one measures the change in absorption of the sample compared to the initial, pre-flash absorption, and collects the data in a matrix, **D**, whose columns are difference spectra taken at consecutive points in time. "Single wavelength" type experiments usually result in more sampling points in the time domain than in the wavelength domain, whereas multichannel spectroscopy leads to more wavelength values than time values. Without loss of generality, this analysis focuses on the latter case, but the conclusions do not in general depend on this choice. Hence, the data matrix, **D**, has dimensions $m \times n$, with $m > n$. According to the Beer−Lambert law the absorption in a sample mixture is the linear combination of the individual component absorption values with their concentration as combination coefficients. Applying this on the mixture *difference* spectra in matrix form, one obtains

$$\mathbf{D} = \epsilon \mathbf{c}^{\mathrm{T}} = \sum_{j=1}^{k} |\epsilon_j\rangle\langle \mathbf{c}_j| \qquad (1)$$

where $|\epsilon_j\rangle$, a column vector of dimensions $m \times 1$, is the *difference* spectrum of the $j$th photocycle intermediate versus the initial state, $\langle \mathbf{c}_j|$, a row vector of dimensions $1 \times n$, is its time dependent concentration, and $\mathbf{c}^{\mathrm{T}}$ is the transpose of **c**. This diadic decomposition shows that the rank of **D**, $\rho(\mathbf{D})$ is at most $k$. The primary information sought is the matrices $\epsilon$ and **c**, where the initial total intermediate concentration, $c_{11} + c_{12} + ... + c_{1k}$ is supposed to yield PCR $\times$ $c$, with $c$ designating the sample concentration. For the sake of simplicity in the following, I will assume that this sum equals 1 (i.e., **c** and $\epsilon$ are properly normalized simultaneously).

The intermediate spectra contain information about the conformation of the chromoprotein, including retinal isomeric state, charge distribution around the retinal, protonation states of the retinal Schiff base (the covalent link binding the retinal to lysine 216) and various amino acid residues, and geometric structure and hydration of the protein. However, the visible absorption spectrum is usually not structured enough to obtain all of this information. Nevertheless, in the process of dissecting the data matrix according to eq 1, the resulting intermediate spectra serve as quality criteria. It is difficult to judge the quality of difference spectra, but once the PCR is determined, the corresponding absolute spectra are readily calculated as $|\mathbf{A}_j\rangle = |\epsilon_j\rangle/\mathrm{PCR} + |\mathbf{A}_0\rangle$, where $|\mathbf{A}_0\rangle$ is the known absorption spectrum of the sample in the initial state. Criteria for the *absolute* spectra can be formulated besides the obvious nonnegativity constraint. These criteria can range from visual appearance to the fit of various nomograms.[12]

The information in the concentration matrix is related to the photocycle scheme, the most sought for target of the analysis. The photocycle scheme includes the intermediates appearing more-or-less sequentially, the transitions between the intermediates, and the rate constants of these transitions. The number of intermediates, $k$, is a priori also unknown. A photocycle scheme can be represented by the $k \times k$ matrix of rate constants **K**, where $K_{ij}$ is the rate constant of the transition from intermediate $j$ to intermediate $i$, $K_{ii}$ is the negative sum of all rate constants leading from intermediate $i$, and the requirement of detailed balance must be fulfilled.[31] Only those photocycle schemes are acceptable, which result in concentration and absorption matrices **c** and $\epsilon$, so that their product reproduces the data matrix **D** within experimental error, the intermediate absorption spectra have adequate spectral properties, and the dependence of the rates on various experimental conditions, most notably temperature, is meaningful. Since eq 1 represents a mathematically underdetermined problem, finding a photocycle scheme from noisy experimental data with all above criteria fulfilled is a difficult task. In the following, some mathematical properties of the decomposition in eq 1 are discussed.

**Problem of Rank Deficiency.** When there are $k$ distinct intermediates in the photocycle, matrices $\epsilon$ and **c** contain $k$ columns. Four cases can be distinguished based on the rank of these matrices.

(i) Both the intermediate difference spectra in matrix $\epsilon$ and the intermediate concentrations in matrix **c** are linearly independent. The rank of both matrices is then $k$. As **c** is full rank, the right pseudoinverse of $\mathbf{c}^{\mathrm{T}}$ exists: $\mathbf{c}^{\mathrm{T}+} = \mathbf{c}(\mathbf{c}^{\mathrm{T}}\mathbf{c})^{-1}$. Since matrix multiplication does not increase the rank of the multiplied matrices, eqs 2 show that the rank of **D** in this case equals $k$

$$\rho(\mathbf{D}) = \rho(\epsilon \mathrm{c}^{\mathrm{T}}) \le \rho(\epsilon), \quad \rho(\epsilon) = \rho(\epsilon \mathrm{c}^{T}\mathbf{c}^{\mathrm{T}+}) \le \rho(\epsilon \mathrm{c}^{T}),$$
$$\text{therefore } \rho(\mathbf{D}) = \rho(\epsilon) \qquad (2)$$

As long as the recovery of the initial state has not started, the total intermediate concentration remains constant, or unity with appropriate normalization. More generally, the total intermediate concentration is one minus the concentration of the recovered initial state

$$\mathbf{c}|\mathbf{e}_k\rangle = |\mathbf{e}_n\rangle - |\mathbf{c}_0\rangle \equiv |\mathbf{b}_n\rangle \qquad (3)$$

where $|\mathbf{e}_k\rangle$ and $|\mathbf{e}_n\rangle$ are all unity column vectors of length $k$ and $n$, respectively, $|\mathbf{b}_n\rangle$ is the stoichiometric vector equal to the total intermediate concentration, and $|\mathbf{c}_0\rangle$ is the concentration of the recovered initial state. If the analysis is restricted to the time interval before initial state recovery, the simpler eq 4 holds

$$\mathbf{c}|\mathbf{e}_k\rangle = |\mathbf{e}_n\rangle \qquad (4)$$

(ii) Matrix $\epsilon$ is full rank, but matrix **c** is not. In practice, this happens for example when two intermediates with distinct spectra are in (infinitely) rapid equilibrium so that their kinetics are strongly coupled. The rank deficiency of **c** can be represented by the equation

$$|\mathbf{c}_k\rangle = \sum_{j=1}^{k-1} \alpha_j |\mathbf{c}_j\rangle \qquad (5)$$

without loss of generality, since the diadic decomposition in eq 1 shows that spectral and kinetic vectors can be simultaneously exchanged (renumbered) in matrices $\epsilon$ and **c**. Hence

$$\mathbf{D} = \epsilon \mathrm{c}^{\mathrm{T}} = \sum_{j=1}^{k-1} (|\epsilon_j\rangle + \alpha_j|\epsilon_k\rangle)\langle \mathbf{c}_j| = \sum_{j=1}^{k-1} \frac{|\epsilon_j\rangle + \alpha_j|\epsilon_k\rangle}{(1+\alpha_j)}\langle \mathbf{c}_j|(1+\alpha_j) = \epsilon' \mathbf{c}'^{\mathrm{T}} \quad (6)$$

The factors $1/(1 + \alpha_j)$ ensure that the mixture difference spectra correspond to unit concentration. This decreases the rank of the data matrix by one, and no analysis will be able to resolve the true intermediate spectra and kinetics with mathematical rigor. All intermediate spectra whose kinetics contaminate the kinetics of the *k*th intermediate will turn out to be mixture spectra with linearly independent kinetics. Rank determination of the data matrix by SVD will yield $k - 1$. Global multiexponential fit to the data will yield a number of exponential components less than the true number of intermediates. Direct model fits to the **D** matrix in this case cannot resolve the true concentration matrix either. On one hand, the investigator ought to anticipate a model with more (how many more?) intermediates than justified by the multiexponential fit. On the other hand, even if the model accurately described all *k* intermediates and their interrelations, the pseudoinverse of the true concentration matrix **c** with the correct rate constants would be close to singular. The iteration to calculate the decomposition in eq 1 involves the multiplication by the pseudoinverse of **c**, resulting in unreliable, erroneous intermediate spectra, which may be carried over to the next iteration step, depending on the algorithm. A simpler model with $k - 1$ intermediates may fit the data well, but at a price to abandon the criteria established for the spectral shape of the intermediates.

The stoichiometric relationships (3) or (4) are still valid for $\mathbf{c}'$

$$\mathbf{c}'|\mathbf{e}_{k-1}\rangle = \sum_{j=1}^{k-1} |\mathbf{c}_j\rangle(1+\alpha_j) = |\mathbf{b}_n\rangle \text{ or } = |\mathbf{e}_n\rangle \quad (7)$$

(iii) Matrix **c** is full rank, but matrix $\epsilon$ is not. Then, at least one of the intermediates has a difference spectrum which is a linear combination of the other spectra

$$|\epsilon_k\rangle = \sum_{j=1}^{k-1} \alpha_j |\epsilon_j\rangle \quad (8)$$

For eq 8 to hold, the *absolute* spectra must fulfill eq 9

$$|A_k\rangle = \sum_{j=1}^{k-1} \alpha_j |A_j\rangle + |A_0\rangle(1 - \sum_{j=1}^{k-1} \alpha_j) \quad (9)$$

From eq 1

$$\mathbf{D} = \epsilon \mathrm{c}^{\mathrm{T}} = \sum_{j=1}^{k-1} |\epsilon_j\rangle(\langle \mathbf{c}_j| + \alpha_j \langle \mathbf{c}_k|) = \epsilon_{k-1} \mathbf{c}'^{\mathrm{T}} \quad (10)$$

Again, the rank of the data matrix decreases by one. The spectra of the intermediates are not mixed, so spectral criteria still work. The inversion of eq 1 is possible when the correct concentration matrix, **c** is used, and accurate intermediate spectra may be obtained. However, there arise limitations in the applicability of the stoichiometric relationships (3) or (4)

$$\mathbf{c}'|\mathbf{e}_{k-1}\rangle = \mathbf{c}|\mathbf{e}_k\rangle + |\mathbf{c}_k\rangle(\sum_{j=1}^{k-1} \alpha_j - 1) \quad (11)$$

The condition for this expression to be equal to $|\mathbf{b}_n\rangle$ or $|\mathbf{e}_n\rangle$ is

$$\sum_{j=1}^{k-1} \alpha_j = 1 \quad (12)$$

Since the variation of the chromoprotein absolute spectrum from intermediate to intermediate is characterized by a shift of the maximum wavelength with amplitude change and potentially some variation in the half width of the main band, distinct absolute spectra are not expected to be composable from the spectra of other intermediates and the initial state. Therefore, true linear dependency of the intermediate difference spectra is not expected either, except for the trivial case of substates with indistinguishable spectra. In the latter case, eq 12 holds, since if, for example, $|\epsilon_k\rangle = |\epsilon_{k-1}\rangle$, then $\alpha_j = \delta_{\mathrm{k-1,j}}$, where $\delta_{\mathrm{ij}}$ is the Kronecker symbol. It is straightforward to show that eq 12 is fulfilled also when more than two intermediates have the same spectrum. The practical consequence is that the bacteriorhodopsin photocycle with at least two, potentially more M intermediates[32] (whose common characteristics are the deprotonated Schiff base) can be analyzed by taking into account the stoichiometric constraints (3) or (4) despite the rank reduction of the data matrix.

(iv) If both matrices $\epsilon$ and **c** are rank deficient, a combination of the difficulties listed under (ii) and (iii) are expected, which may make the correct natural decomposition of the data matrix mathematically impossible.

In the following, it is assumed that both matrices $\epsilon$ and **c** are full rank, or **c** is full rank and $\epsilon$ is not, but only because some intermediate spectra are identical. As we have seen in eq 10, the latter case reduces to the case of full rank matrices. Hence the rank of the relevant matrices is

$$\rho(\mathbf{D}) = \rho(\epsilon) = \rho(\mathbf{c}) \equiv r \quad (13)$$

**Relation between the Natural Decomposition and SVD.** Singular value decomposition results in 3 matrices of special properties[33]

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}},\ \mathbf{U}_{m\times n},\ \mathbf{V}_{n\times n},\ \mathbf{S}_{n\times n}\text{: diagonal,}\quad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{1},\ \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{1},\ \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{1} \quad (14)$$

where **1** represents the unit matrix, and the elements of **S**, the singular values, are the positive square roots of the eigenvalues of the square matrix $\mathbf{D}^{\mathrm{T}}\mathbf{D}$. If **D** is a noise free matrix of rank $r < n$, then only the first *r* singular values are nonzero; therefore, the first *r* columns of **U** and **V**, and the upper left $r \times r$ submatrix of **S** reconstruct the data matrix. For noisy data the effective rank can be determined by a combination of criteria.[33,34] First, $\sum_{i=r+1}^{n} \mathbf{S}_{ii}^2$ should be less than the average noise of **D**, estimated as $n \times m \times \sigma^2$, where $\sigma^2$ is the average variance of the elements of **D**. Second, the autocorrelations of the first *r* columns of **U** and **V**, the spectral and kinetic eigenvectors, should be high enough, which distinguishes them from those eigenvectors carrying only noise. The rotation algorithm of Henry and Hofrichter[33] is a recommended tool at this stage to "pull forward" signals with high autocorrelation which are contained in columns of **U** or **V** otherwise discarded based on the noise criterion. It is assumed in the following that the rank has been determined based on these criteria, and the dimension of **U**, **V**, and **S** in eq 14 are $m \times r$, $n \times r$, and $r \times r$, respectively:

$$\mathbf{D} \cong \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}},\ \mathbf{U}_{m\times r},\ \mathbf{V}_{n\times r},\ \mathbf{S}_{r\times r}\text{: diagonal,}\quad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{1}_{r\times r},\ \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{1}_{r\times r},\ \text{but } \mathbf{V}\mathbf{V}^{\mathrm{T}} \neq \mathbf{1} \quad (15)$$

where **D** now means the reconstructed, noise filtered data matrix of rank *r*.

The columns of $\epsilon$ are linear combinations of the columns of **U**. This is the consequence of **c** being full rank, and therefore the existence of the right pseudoinverse $\mathbf{c}^{T+}$

$$\epsilon = \epsilon\mathbf{c}^T\mathbf{c}^{T+} = \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{c}^{T+} \equiv \mathbf{U}\mathbf{R}, \text{ and } \mathbf{R} = \mathbf{U}^T\epsilon \quad (16)$$

Equation 16 allows the definition of the combination coefficients of the pure intermediate spectra

$$\epsilon = \mathbf{U}\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{S}^{-1}\mathbf{R} \equiv \mathbf{U}\mathbf{S}\mathbf{V}_\epsilon^{\mathrm{T}}, \text{ where } \mathbf{V}_\epsilon = \mathbf{R}^{\mathrm{T}}\mathbf{S}^{-1} \quad (17)$$

and we utilized the fact that the diagonal matrix **S** with nonzero diagonal elements is invertable.

The columns of **U** are linear combinations of the columns of $\epsilon$. Multiplying the equation $\epsilon\mathbf{c}^{\mathrm{T}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$ by **V** and $\mathbf{S}^{-1}$ from the right yields

$$\epsilon\mathbf{c}^{\mathrm{T}}\mathbf{V}\mathbf{S}^{-1} = \mathbf{U} \equiv \epsilon\mathbf{R}', \text{ and } \mathbf{R}' = \mathbf{c}^{\mathrm{T}}\mathbf{V}\mathbf{S}^{-1} \quad (18)$$

The relationship between **R** and **R′** is obtained as follows:

$$\mathbf{U} = \epsilon\mathbf{R}' = \mathbf{U}\mathbf{R}\mathbf{R}', \text{ which, when left multiplied by } \mathbf{U}^{\mathrm{T}}, \text{ yields } \mathbf{R}\mathbf{R}' = \mathbf{1} \quad (19)$$

**R**, which is a $r \times r$ square matrix, is invertable if $\rho(\mathbf{R}) = r$. Let us assume that $\rho(\mathbf{R}) = r - 1$. Then

$$|\mathbf{R}_r^{\mathrm{T}}\rangle = \sum_{i=1}^{r-1}\alpha_i|\mathbf{R}_i^{\mathrm{T}}\rangle, \epsilon = \mathbf{U}(\mathbf{R}^{\mathrm{T}})^{\mathrm{T}} = \sum_{i=1}^{r-1}(|\mathbf{U}_i\rangle + \alpha_i|\mathbf{U}_r\rangle)\langle\mathbf{R}_i^{\mathrm{T}}| \quad (20)$$

which means that $\epsilon$ can be decomposed into the sum of $r - 1$ diades. This contradicts the assumption that $\rho(\epsilon) = r$. It is shown, therefore, that $\mathbf{R}^{-1}$ exists, and hence, from eq 19, $\mathbf{R}' = \mathbf{R}^{-1}$.

The columns of **V** are linear combinations of the columns of **c**. Multiplying the equation $\epsilon\mathbf{c}^{\mathrm{T}} = \mathbf{U}\mathbf{R}\mathbf{c}^{\mathrm{T}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$ by $\mathbf{U}^{\mathrm{T}}$ and $\mathbf{S}^{-1}$ from the left yields

$$\mathbf{S}^{-1}\mathbf{R}\mathbf{c}^{\mathrm{T}} = \mathbf{V}^{\mathrm{T}}, \quad \mathbf{V} = \mathbf{c}\mathbf{R}^{\mathrm{T}}\mathbf{S}^{-1} = \mathbf{c}\mathbf{V}_\epsilon \quad (21)$$

The columns of **c** are linear combinations of the columns of **V**. This follows from eq 21 and the existence of the inverse of **R**, which also ensures the invertability of $\mathbf{V}_\epsilon$

$$\mathbf{c} = \mathbf{V}\mathbf{S}(\mathbf{R}^{\mathrm{T}})^{-1} = \mathbf{V}\mathbf{V}_\epsilon^{-1} \quad (22)$$

An important consequence of the intimate link between matrices **V** and **c** and the linear independence of the exponential functions is that, if the photocycle processes consist of first order (or pseudo first order) reactions, as generally assumed, and are therefore satisfactorily described by the rate constant matrix **K**, the columns of **V** will be linear combinations of the same exponentials as the columns of **c**. These phenomenological rate constants can therefore be obtained in principle from a global multiexponential fit to the **V** matrix.

**Stoichiometric Equation.** From the assumptions on the rank of $\epsilon$ and **c**, the validity of eqs 3 or 4 follows. Substituting eq 22 into eq 3, we obtain

$$\mathbf{V}\mathbf{V}_\epsilon^{-1}|\mathbf{e}_r\rangle = |\mathbf{b}_n\rangle, \mathbf{V}_\epsilon^{-1}|\mathbf{e}_r\rangle = \mathbf{V}^{\mathrm{T}}|\mathbf{b}_n\rangle, \mathbf{V}\mathbf{V}^{\mathrm{T}}|\mathbf{b}_n\rangle = |\mathbf{b}_n\rangle, \text{ or } \mathbf{V}\mathbf{V}^{\mathrm{T}}(\mathbf{c}|\mathbf{e}_r\rangle) = \mathbf{c}|\mathbf{e}_r\rangle \quad (23)$$

This shows that the $n \times 1$ element stoichiometric vector $|\mathbf{b}_n\rangle$, which equals the sum of the intermediate concentrations, is an eigenvector of the square matrix $\mathbf{V}\mathbf{V}^{\mathrm{T}}$ with eigenvalue 1. Recasting eq 23 yields eq 24

$$\mathbf{V}|\mathbf{q}_r\rangle = |\mathbf{b}_n\rangle, \text{ where } |\mathbf{q}_r\rangle = \mathbf{V}^{\mathrm{T}}|\mathbf{b}_n\rangle \quad (24)$$

What is the corresponding stoichiometric equation for matrix $\mathbf{V}_\epsilon$, the combination coefficients of the pure intermediate spectra? Combining eqs 23 and 24 yields

$$\mathbf{V}_\epsilon^{-1}|\mathbf{e}_r\rangle = |\mathbf{q}_r\rangle, \text{ therefore } \mathbf{V}_\epsilon|\mathbf{q}_r\rangle = |\mathbf{e}_r\rangle \quad (25)$$

The first half of eq 24 is valid even for a truncated stoichiometric vector $|\mathbf{b}_t\rangle$ and corresponding $\mathbf{V}_{\mathrm{t}}$ matrix where, say, only the first *t* time points are considered. The second half, which is the derivation of the $|\mathbf{q}_r\rangle$ parameter vector, is not, since it relies on the orthonormality of the full column vectors in **V**. A stoichiometric equation for the truncated data set can now be obtained by a least squares solution for $|\mathbf{q}_r\rangle$ of eq 26, provided that $|\mathbf{b}_t\rangle$ is known and the left pseudoinverse $\mathbf{V}_{\mathrm{t}}^{+}$ exists

$$\mathbf{V}_t|\mathbf{q}_r\rangle = |\mathbf{b}_t\rangle, \text{ where } |\mathbf{q}_r\rangle = \mathbf{V}_t^{+}|\mathbf{b}_t\rangle, \mathbf{V}_t^{+} = (\mathbf{V}_t^T\mathbf{V}_t)^{-1}\mathbf{V}_t^{\mathrm{T}} \quad (26)$$

A rank analysis of $\mathbf{V}_t$ is in order here, because, if its rank is less than *r*, a new SVD on the truncated data matrix must be performed, and the process repeated with the reduced rank. In case of noisy data, this rank analysis yields only an effective rank, and its reliability depends on the amplitude of the noise. From the relation defining the parameters of the stoichiometric equation, $|\mathbf{q}_r\rangle = \mathbf{V}^T|\mathbf{b}_n\rangle$, it follows that the elements of $|\mathbf{q}_r\rangle$ will tend to zero with increasing estimated rank, because the corresponding columns of **V** will carry noise of random amplitudes and sign, and the elements of $|\mathbf{b}_n\rangle$ are between 0 and 1. Therefore, an overestimate of *r* is less dangerous than its underestimate.

In practice $|\mathbf{b}_n\rangle$ is not known. One can assume, however, that up to a certain time during the photocycle the recovery of the initial state is negligible, so that $|\mathbf{b}_t\rangle = |\mathbf{e}_t\rangle$, therefore

$$\mathbf{V}_t|\mathbf{q}_r\rangle = |\mathbf{e}_t\rangle \quad (27)$$

This equation is sometimes referred to as the equation of the stoichiometric plane (or stoichiometric hypersurface), since its form resembles the equation of a plane in 3D space.

The stoichiometric eq 27 is a consequence of the stoichiometric relation $\mathbf{c}_t|\mathbf{e}_r\rangle = |\mathbf{e}_t\rangle$ for the first *t* time points. If there is no acceptable solution to eq 27, i.e., there exists no linear combination of the columns of **V** which would approximate the all unity vector $|\mathbf{e}_t\rangle$ sufficiently, then the stoichiometric relation $\mathbf{c}_t|\mathbf{e}_r\rangle = |\mathbf{e}_t\rangle$ cannot hold either. Is, however, the existence of the stoichiometric plane (eq 27) sufficient for the existence of the stoichiometric relationship?

Suppose that one finds a solution to eq 27, i.e., a $|\mathbf{q}_r\rangle$ vector which, when multiplied by $\mathbf{V}_t$, yields $|\mathbf{e}_t\rangle$ with acceptable fluctuations, if any. Consider any real vector $|\mathbf{p}_r\rangle$ and the linear combination $|\mathbf{b}_t\rangle = \mathbf{V}_t|\mathbf{p}_r\rangle$. Let us find any invertable $r \times r$ matrix $(\mathbf{V}_\epsilon^*)^{-1}$ with the property $(\mathbf{V}_\epsilon^*)^{-1}|\mathbf{e}_r\rangle = |\mathbf{p}_r\rangle$. An infinite number of such matrices should exist, and one example is a diagonal matrix with $p_1, p_2, \ldots, p_r$ in its main diagonal. This matrix will generate a potential concentration matrix with the properties

$$\mathbf{c}^* = \mathbf{V}_t(\mathbf{V}_\epsilon^*)^{-1}, \mathbf{c}^*|\mathbf{e}_r\rangle = \mathbf{V}_t|\mathbf{p}_r\rangle = |\mathbf{b}_t\rangle \neq |\mathbf{e}_t\rangle \text{ with arbitrary } |\mathbf{p}_r\rangle \quad (28)$$

Mathematically, therefore, it is always possible to generate concentration matrices whose rows, when summarized, will not

yield all ones, despite the fact that the **V** matrix fulfills the stoichiometric equation. The number of potential concentration matrices is drastically reduced due to the physical requirements of nonnegative concentrations and the simultaneous criteria regarding the acceptable intermediate spectra. Nevertheless, the extrapolation from the stoichiometric equation (eq 27) to the stoichiometric relationship (eq 4) remains an assumption without rigorous proof, whose validity can only be justified in retrospect, when the obtained concentration and absorption matrices fulfill the respective criteria.

Equation 25 for the combination coefficients of the pure intermediate spectra is valid only when the true $|\mathbf{q}_r\rangle$ vector is used, i.e., the one which satisfies $\mathbf{c}_t|\mathbf{e}_r\rangle = \mathbf{V}_t|\mathbf{q}_r\rangle$ with the true **c** matrix. If, by accident, there exists at the same time a vector $|\mathbf{q}_r'\rangle$ so that $\mathbf{V}_t|\mathbf{q}_r'\rangle = |\mathbf{e}_t\rangle$, searching for the pure intermediate spectra on the stoichiometric plane defined by $|\mathbf{q}_r'\rangle$ will yield invalid spectra.

**Self-Modeling.** With these caveats in mind, I continue with the description of self-modeling, for which eqs 25 and 27 serve as the basis. In practice, one tries to find the equation for the stoichiometric plane (eq 27) and then find the $\mathbf{V}_\epsilon$ coefficients by moving on the surface of this plane. This geometric analogy helps to visualize the process in up to 3 dimensions.

In 3D, the combination coefficients of the pure intermediates must define the vertexes of a triangle on the stoichiometric plane. To ensure nonnegativity of the concentrations, this triangle must contain all points defined by the rows of the $\mathbf{V}_t$ matrix. Any points outside the triangle, that is, would mean a mixture with at least one of the intermediates having negative concentration. If a vertex of the triangle is actually realized, this means that the sample at the corresponding time was fully converted to one of the intermediates. A point on one of the sides of the triangle would mean a mixture of two intermediates, without any measurable amount of the third one.

Finding the sides and vertexes of the triangle is based on spectral criteria. These criteria relate to the *absolute* spectra, which can only be calculated if one knows the PCR (see above). In the bacteriorhodopsin photocycle the amplitude of the absorption change at any single wavelength (such as 412 or 570 nm) cannot be used directly to estimate the PCR, due to the facts that all intermediates absorb at these wavelengths, and except for special cases such as the Asp96 → Asn mutant protein, there exist unknown mixtures of intermediates at all times during the photocycle. The absorption increase at 412 nm will rather accurately yield the amount of the M intermediate but not the total amount of the intermediates. Estimating the remaining amount of BR, i.e., 1-PCR, from the response to a second laser flash is also inaccurate. There is always the possibility of exciting some of the intermediates present at the time of the second flash, which may result in transitions which are normally not part of the photocycle. In addition, due to the dependence of the photocycle on the amount of BR molecules cycling,[35] the second flash may induce a kinetically different photocycle in the BR molecules initially not excited by the first flash.

The M intermediate has zero absorption at higher than ~540 nm; therefore, the M − BR difference spectrum should exactly follow the shape of the −BR spectrum in this range. This allows the PCR to be estimated from multichannel data by the method of target testing, provided that the photocycle involves an M intermediate, and it is not in a very rapidly evolving equilibrium with other intermediates, so that its spectrum is separable from the spectra of all other intermediates. Experience shows that this is the case for wild type and most mutant bacteriorhodopsin molecules.

Suppose that we can separate a time interval where there is clear indication from the raw difference spectra of the presence of M. After performing SVD and rank determination, eq 27 is solved for the parameters of the stoichiometric plane. If a solution is found, we assume that this reflects the fact that no BR recovery has yet taken place. If we have a target difference spectrum for M-BR, $|\epsilon_M^t\rangle$, whose amplitude corresponds to full conversion to M, we assume that this differs from the actual M-BR spectrum only in its amplitude, and therefore

$$|\epsilon_M^t\rangle = p|\epsilon_M\rangle = \mathbf{U}\mathbf{S}p\langle\mathbf{V}_M|^{\mathrm{T}} \text{ therefore } \mathbf{S}^{-1}\mathbf{U}^T|\epsilon_M^t\rangle = p\langle\mathbf{V}_M|^{\mathrm{T}} \tag{29}$$

where p is the unknown scaling factor. Substituting this into eq 27 we obtain

$$p\langle\mathbf{V}_M|\mathbf{q}_r\rangle = p \tag{30}$$

Equation 30 shows that $p$ equals 1/PCR. The target spectrum may be imported from another experiment or, if this is not reliable, one can use the −BR spectrum usually available from a steady-state measurement, and restrict eq 29 to the wavelength range, say, $\lambda > 540$ nm

$$|-\mathbf{BR}_\lambda\rangle = p|\epsilon_{M,\lambda}\rangle = \mathbf{U}_\lambda\mathbf{S}p\langle\mathbf{V}_M|^{\mathrm{T}} \tag{31}$$

where the $\lambda$ subscripts indicate that the corresponding spectra are truncated. Equation 31 should now be solved for the $p\langle\mathbf{V}_M|$ scaled combination coefficients by a linear least-squares fit or, in other words, by premultiplying with the pseudoinverse of $\mathbf{U}_\lambda\mathbf{S}$. Then, PCR is obtained from eq 30, and the full $|\epsilon_M\rangle$ difference spectrum is calculated with the combination coefficients $\langle\mathbf{V}_M|$. This procedure yields, therefore, the PCR, the difference (and absolute) spectrum of M, and the vertex of M on the stoichiometric plane. The quality of the M *absolute* spectrum and, specifically, the requirement that it be zero above 540 nm, is checked at this stage.

The search for the vertexes corresponding to other intermediates is facilitated by removing the remaining random variations from the **V** matrix. The columns of **V** can be subjected to a global multiexponential fit, and the resulting smooth fitted functions can replace the **V** matrix. This method, termed here as exponential fit assisted self-modeling (SVD-EFASM), is useful to identify trends of evolution of the photocycle from intermediate to intermediate when plotting the columns of the smooth **V** as a function of each other. These plots will exhibit regions where consecutive points fall almost on straight lines and sometimes sharp turns between these regions. Unfortunately, the "lines" do not usually coincide with the edges of the stoichiometric polyhedron, nor do the turns, or their extrapolation as the crossing points of the "lines", coincide with the vertexes representing pure intermediates. This is the consequence of the temporal overlap of the intermediates resulting in mixtures at all times, usually consisting of more than two intermediates. Nevertheless, these plots can guide the search for the true vertexes by indicating trends. These trends in **V** are followed during self-modeling, and the resulting spectra calculated each time, until certain spectral criteria are fulfilled, indicating that a side of the triangle, or a vertex, is reached. Another advantage of EFASM is that it allows a $t = 0$ extrapolation of the **V** matrix, and the resulting combination coefficients may yield the first

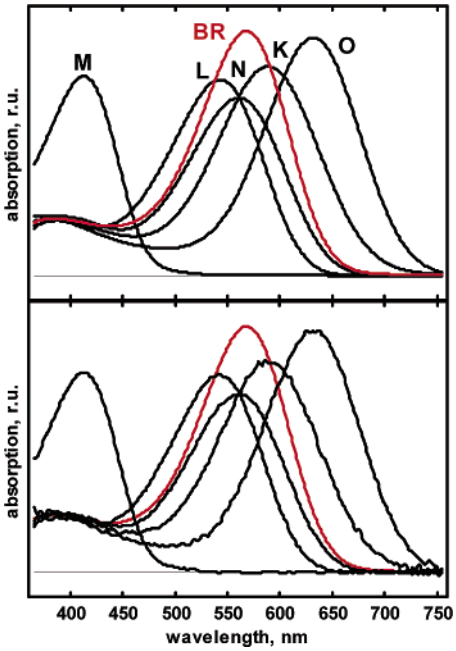


**Figure 1.** Input initial state BR spectrum (red) and intermediate spectra (black) for the simulated photocycle data (top). Output spectra of the intermediates obtained by SVD-EFASM (bottom).

intermediate resolvable by the actual experiment, usually the K or KL intermediate.

It is commonly assumed that the first half of the photocycle consists of consecutive K, L, and M intermediates, allowing substates of M and, perhaps, of L. Once the $t = 0$ extrapolation yields K, this vertex can be connected to a later point, and moving along this line should lead to a point where all contribution by K is removed. This point should consequently lie on the LM side of the triangle. The spectral criterion here is an absolute spectrum which decays to the baseline at wavelengths where neither L nor M is expected to absorb, say, above 690 nm. Since the M vertex is also known, two points now define this line, and along it one can find the L vertex in what is actually a 2D self-modeling. The spectral criterion is the average absorption of L to be either equal to, or close to the average absorption of BR in the range of the minor band close to 400 nm. There is a certain arbitrariness in the selection of the average absorption of L in the blue region, which is carried over to the resulting calculated kinetics of L and M.

In higher dimensions, it is still possible to cancel the contribution of an intermediate to a mixture difference spectrum by connecting the intermediate's vertex to the point representing the mixture, and moving along this line beyond the mixture point until some spectral criterion indicates total removal of the intermediate considered, provided that such spectral criterion can be found. This procedure automatically guarantees that the calculated spectrum's amplitude will be corrected for the removal of one of its components, since the combination coefficients remain on the stoichiometric hypersurface. Several such points will now define a new stoichiometric surface with a dimensionality decreased by one, and this new surface could in principle be the starting point of a new self-modeling procedure. This approach was successfully applied to the second half of the Glu204 → Gln mutant protein's photocycle,[25] where removal of the contribution by M resulted in a 3 component system with the L, N, and O intermediates participating.

The analysis in ref 25 was meant to demonstrate that it was possible to extract acceptable intermediate spectra from the data matrix with acceptable kinetics but did not include a thorough search for the best reaction scheme. This demonstration was important in view of the fact that several investigators have offered solutions to the photocycle of bacteriorhodopsin[36,37] as well as to the analogous reactions of rhodopsin and sensory rhodopsin[20,21] with multicolored intermediate absorption bands or major shoulders.

**TABLE 1: Input Time Constants of the Simulated Photocycle**

| transition | $\tau$, $\mu$s | transition | $\tau$, $\mu$s | transition | $\tau$, ms |
|---|---|---|---|---|---|
| $K \rightarrow L_1$ | 1 | $L_2 \leftarrow M_1$ | 100 | $N_1 \rightarrow N_2$ | 10 |
| $K \leftarrow L_1$ | 2 | $M_1 \rightarrow M_2$ | 60 | $N_1 \leftarrow N_2$ | 30 |
| $L_1 \rightarrow L_2$ | 10 | $M_1 \leftarrow M_2$ | 1000 | $N_2 \rightarrow O$ | 20 |
| $L_1 \leftarrow L_2$ | 30 | $M_2 \rightarrow N_1$ | 1000 | $N_2 \leftarrow O$ | 500 |
| $L_2 \rightarrow M_1$ | 30 | $M_2 \leftarrow N_1$ | 1000 | $O \rightarrow BR$ | 50 |

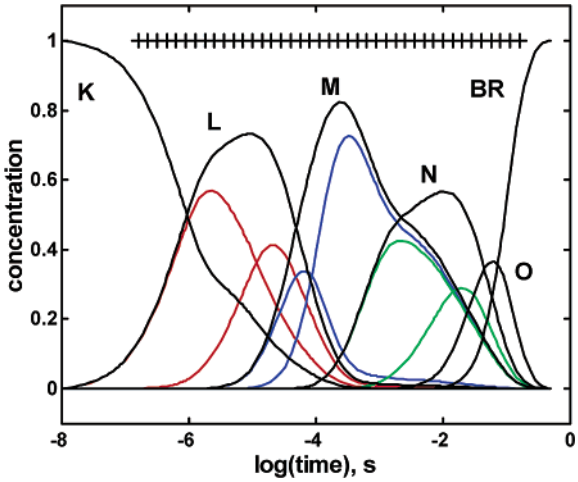


**Figure 2.** Simulated input kinetics of the photocycle intermediates. Red, blue, and green lines are the kinetics of the substates of L, M, and N, respectively. The vertical lines on top mark the 41 sampling times for the creation of the input concentration matrix.

## Analysis of Simulated Data

**Construction of Realistic Simulated Data.** A data matrix of difference spectra was created from input intermediate absolute spectra and kinetics and from a noise matrix similar to the noise pattern usually met in my (and other investigators') optical multichannel experiments. The smooth absolute spectra were obtained by fitting the pH averaged intermediate spectra found by a Monte Carlo search method[5] with the nomogram in[12] (Figure 1, top).

The linear, reversible photocycle scheme consisting of 5 spectrally and 8 kinetically distinct intermediates $K \rightleftarrows L_1 \rightleftarrows L_2 \rightleftarrows M_1 \rightleftarrows M_2 \rightleftarrows N_1 \rightleftarrows N_2 \rightleftarrows O \rightarrow BR$ was considered, in accordance with reported substates of the L, M, and N intermediates under various experimental conditions.[38−41] The corresponding set of differential equations was integrated with the Runge Kutta method (program SIMULATE written in Microsoft BASIC PDS, all other calculations were performed in MATLAB, The MathWorks, Natick, MA) with input time constants listed in Table 1. The resulting intermediate concentrations are shown in Figure 2.

Subsequently, 41 logarithmically equidistant time points were selected for sampling the kinetics of the five intermediates (i.e. K, $L_1 + L_2$, $M_1 + M_2$, $N_1 + N_2$, and O) to yield the simulated input concentration matrix. This matrix was multiplied by the difference spectrum matrix obtained from the input absolute spectra and a PCR value of 0.3, resulting in the noise free data matrix.

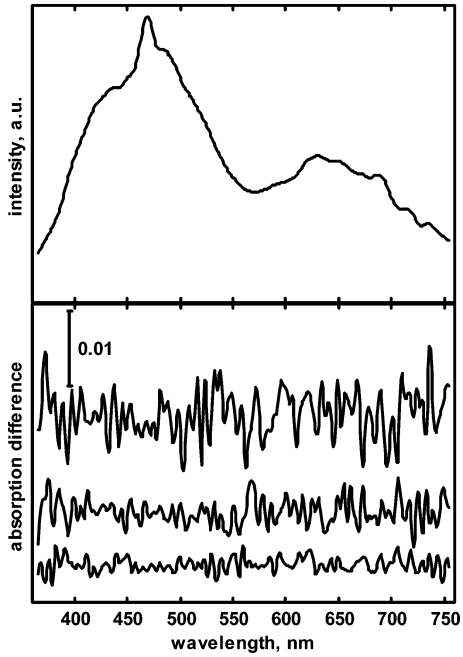


**Figure 3.** Top: Typical light intensity spectrum of a high pressure Xe lamp, measured by the optical multichannel analyzer behind a sample of wild type bacteriorhodopsin. The square root of this intensity spectrum is expected to be proportional to the amplitude of the wavelength dependent noise spectrum. Bottom: Simulated noise spectra for 3 time domains of increasingly longer detector gate pulses (see text). The dependence of the noise amplitude on the light intensity spectrum is also taken into account.

Noise spectra were constructed by generating Gaussian distributed random numbers over a wavelength scale of 3 nm step size and converting the spectra to the final 1 nm step size scale using the spline function. This procedure mimicks the crosstalk of neighboring diodes of the array detector, which results in experimental noise patterns of short distance autocorrelation.

The noise spectra were then divided point-by-point by the square root of a typical light intensity spectrum of a high-pressure xenon lamp, measured behind a sample of wild-type bacteriorhodopsin (Figure 3, top), since the experimental noise amplitude is proportional to the square root of the number of accumulated photons.

In addition, usual measurements over many decades in time are split into time domains of increasing gate pulse width of the diode array detector, to enable better signal-to-noise where long delay times allow longer integration times (i.e., more accumulated photons). This was also taken into account in the simulation by multiplying the noise spectra by 3 and 2 in the time intervals 150 ns to 1.5 $\mu$s and 1.5−15 $\mu$s, respectively. Three typical simulated noise spectra for the three time domains are shown in Figure 3, bottom. In a real experiment, such noise spectra can be measured under conditions otherwise identical to the flash photolysis measurements by blocking the actinic laser light before the sample. The variance of the noise spectrum for the shortest time domain in Figure 3 is 0.54% of the variance of the first difference spectrum, a fraction typically observed in good quality time-resolved spectra on WT BR.

The so constructed spectrally and temporally heterogeneous noise matrix was finally added to the noise free simulated data matrix to yield the input noisy data matrix shown in Figure 4. The following analysis uses this matrix as well as the BR absolute spectrum (Figure 1 top, red spectrum), the light intensity spectrum (Figure 3, top) and the three noise spectra

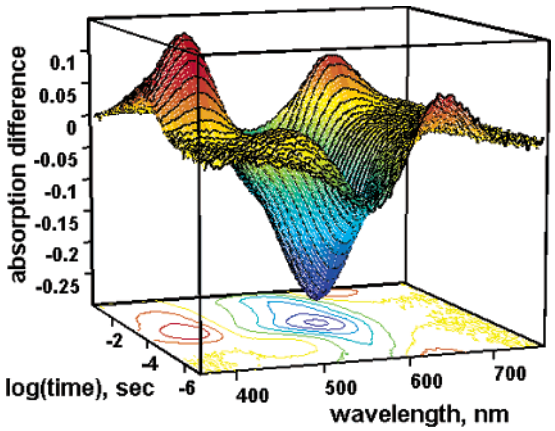


**Figure 4.** Plot of the simulated input noisy difference spectrum matrix.

**TABLE 2: Singular Values and Autocorrelations of the Spectral and Temporal Eigenvectors Calculated from the Data Matrix Normalized to Uniform Noise[a]**

| no. of component, $j$ | singular value, $\mathbf{S}_{jj}$ | $\sum_{i=j}^{41} \mathbf{S}_{i,i}^2$ | autocorrelation[b] of **U** column $j$ | autocorrelation of **V** column $j$ |
|---|---|---|---|---|
| 1 | 6.060 | 38.8 | 0.986 | 0.994 |
| 2 | 1.181 | 2.03 | 0.983 | 0.936 |
| 3 | 0.777 | 0.634 | 0.957 | 0.959 |
| 4 | 0.148 | 0.0312 | 0.981 | 0.940 |
| 5 | 0.057 | 0.00924 | 0.876 | 0.848 |
| 6 | 0.020 | 0.00599 | 0.050 | 0.050 |
| 7 | 0.019 | 0.00560 | −0.046 | −0.157 |
| 8 | 0.019 | 0.00523 | −0.006 | −0.387 |

[a] The estimated standard deviation of the noise is 0.00728, falling between the fifth and sixth numbers in column 3. [b] Calculated with a shift of 9 nm to avoid interference by the short distance spectral autocorrelation of the noise.

(Figure 3, bottom) as the sole information available from visible, non polarized optical data to dissect the bacteriorhodopsin photocycle.

**Singular Value Decomposition and Global Multiexponential Fit.** Singular value decomposition works best in determining the rank of the data matrix and reducing the random noise through the reconstruction of the data matrix when the noise is homogeneous.[33] The data matrix was therefore first weighted to yield a new matrix, by considering the light intensity spectrum and the available noise spectra for the three time domains mentioned above. Difference spectra were multiplied by the square root of the intensity spectrum point by point and divided by 3 and 2 in the time domains 150 ns to 1.5 $\mu$s and 1.5−15 $\mu$s, respectively. The weighted data matrix was subjected to SVD. The first eight singular values and the autocorrelations of the corresponding spectral and temporal eigenvectors are listed in Table 2.

The average noise content of the weighted data matrix was obtained by calculating the average standard deviation of the three noise spectra, weighted similarly to the data matrix. This estimated standard deviation was 0.00728. A rank of 5 was unambiguously obtained by both the noise test (Table 2, column 3) and the autocorrelation test, in accordance with the 5 spectrally distinct intermediates. The data matrix was therefore reconstructed using the first 5 SVD components, and the spectral and temporal weightings reversed. This matrix was subjected to SVD again, and the first 5 significant components were stored for further analysis.

Similar treatment was applied to determine the rank of truncated data matrices consisting of the first $k$ columns only. A rank of 3 was obtained up to the 25th spectrum, a rank of 4

**TABLE 3: Goodness of the Global Multiexponential Fits to the 5 Significant Columns of the V Matrix Weighted by the Corresponding Singular Values[a]**

| | standard deviation of the global multiexponential fit | | |
|---|---|---|---|
| no. of exponentials | $(\mathbf{SV}^T)^T$, $\times 10^3$ | original data matrix, $\times 10^3$ | SVD reconstituted data matrix, $\times 10^4$ |
| 8 | 3.12 | 1.52 | 3.54 |
| 7 | 6.01 | 1.63 | 6.81 |
| 6 | 16.2 | 2.37 | 18.4 |
| 5 | 23.4 | 3.05 | 26.6 |

[a] The standard deviation of the fits to the $(\mathbf{SV}^T)^T$ and that of the difference matrices between the exponential reconstruction and the original and the SVD-reconstructed matrices are listed.

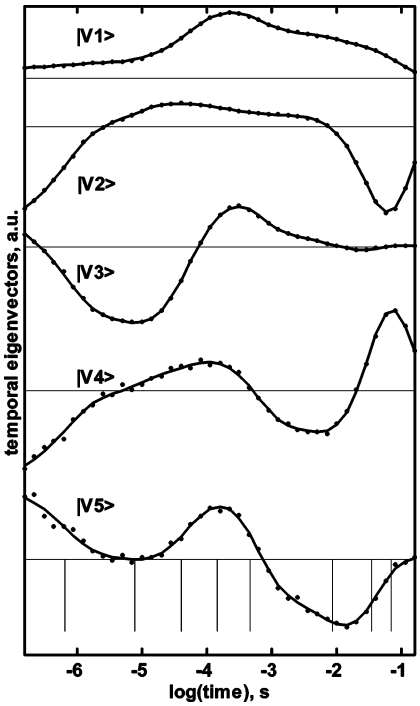


**Figure 5.** Global multiexponential fit (solid lines) to the first 5 significant kinetic eigenvectors of the SVD of the data matrix (dots). Vertical lines in the bottom show the time constants of the 8 phenomenological exponentials obtained by this fit: 666 ns, 7.86 $\mu$s, 40.2 $\mu$s, 145 $\mu$s, 457 $\mu$s, 8.71 ms, 34.3 ms, and 68.2 ms.

up to the 34th spectrum, and a rank of 5 from the first 35 spectra on. It must be noted that at the 25th simulated time point the input amount of N was 0.29, and at the 34th time point, the input amount of O was 0.11. Thus, at the present noise level, SVD tends to underestimate the rank of the data matrix at the onset of a new intermediate. It is safe to assume that the first 19 difference spectra contain only K, L, and M (the input amount of N is < 1%), and the first 30 difference spectra contain only K, L, M, and N (the input amount of O is < 1%), which is equivalent to stepping back in time by more than half but less than 1 order of magnitude.

Global multiexponential fit to the five columns of **V** weighted by the corresponding singular values, i.e., to columns of $(\mathbf{SV}^T)^T$, were performed considering 5, 6, 7, and 8 exponentials. The results in Table 3 and Figure 5 show that 8 exponentials are required to satisfactorily fit the data, in accordance with the input of 8 kinetically distinct intermediates. Increasing the number of exponentials from 5 to 8 improved considerably not only the fit to the kinetic eigenvectors but also to the original and to the SVD reconstructed data matrices. The latter were characterized by the standard deviation of the difference between

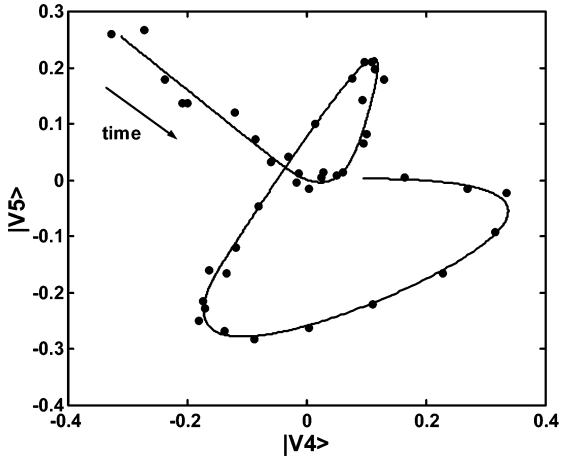


**Figure 6.** Plot of the fifth temporal eigenvector versus the fourth one (dots) and the corresponding pair of fitted multiexponential functions (line).

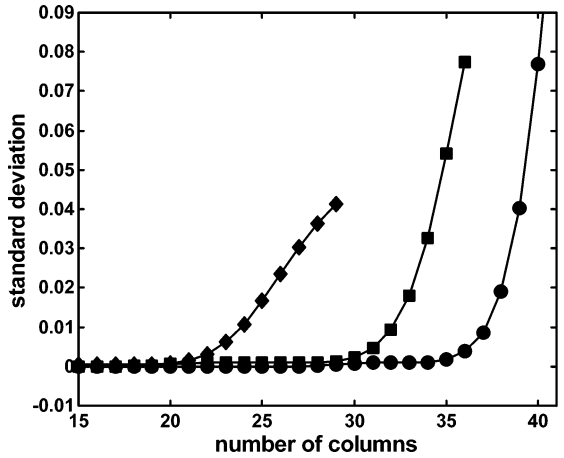


**Figure 7.** Test of the validity of the stoichiometric equation for the simulated data. The standard deviation between the left and right sides of eq 27 is plotted for data matrices consisting of the first 15, 16, ..., 41 columns in rank 3 (diamonds), rank 4 (squares), and rank 5 (circles) approximation.

the respective data matrices and the matrix obtained as the product of the exponential amplitude spectra and kinetics.

The multiexponential fits to the temporal eigenvectors were used in the following instead of the temporal eigenvectors themselves. The advantage of this exponential fit assisted self-modeling (EFASM) is demonstrated in Figure 6, where the fourth and fifth temporal eigenvectors and the corresponding multiexponential fits are plotted (similar plots could be shown for all other pairs of components).

The introduction of the fit removes residual scattering in the temporal eigenvectors and clearly defines smooth lines which characterize the time evolution of the sample. The initial and the turning points must be close to the vertexes corresponding to the pure intermediates K, L, M, N, and O, and the goal of self-modeling is to find these vertexes and calculate the respective pure spectra, as described in the Theoretical section.

**Stoichiometric Equation. Target Testing to Obtain the M Spectrum and PCR.** Truncated data matrices consisting of the first 15, 16, etc. columns of the exponential fit reconstructed data matrix were subjected to SVD, and the validity of the stoichiometric equation (eq 27) was tested, assuming a rank of 3, 4, and 5, by calcualting the standard deviation of the difference $\mathbf{V}_t|\mathbf{q}_r\rangle - |\mathbf{e}_t\rangle$ (Figure 7).

In the rank 3 approximation, the stoichiometric equation failed from the 20th point and in the rank 4 approximation from the

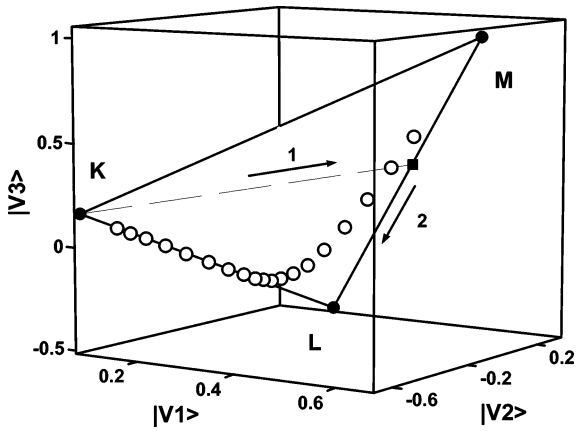


**Figure 8.** Self-modeling to obtain the L spectrum. Open circles represent the rows of the **V** matrix from the rank 3 SVD of the first 19 columns of the data matrix. The vertexes of M and K were obtained by target testing and extrapolation to zero time, respectively. The LM side of the stoichiometric triangle was located by moving along the dashed line connecting the K vertex and the 18th data point until all absorption of the resulting absolute spectrum was removed in the red spectral range, where neither L nor M is expected to absorb (square). The L vertex was found along the LM line where the resulting L spectrum's average absorption equaled that of the BR spectrum in the 390−430 nm range.

30th point. The meaning of these results is not the onset of the recovery of the initial state but the underestimation of the rank. In other words, this treatment signals the increase of rank from 3 to 4 around the 20th time point and that from 4 to 5 around the 30th time point, in a much better agreement with the input data than the noise and autocorrelation analyses described above. In the rank 5 approximation, the breakdown of the stoichiometric equation starts around point 34, correctly reflecting the >1% presence of BR in the input kinetics.

In the adequate rank 3 decomposition of the first 19 columns, it is possible to recover the spectrum of M and the PCR, as described in the Theoretical section (eq 31). The blue end of the target region for the −BR spectrum was varied from 480 to 580 nm, and the best M absolute spectra were obtained, as expected, when the blue end of the target region was around 540 nm. The quality criterion for M was its standard deviation from the baseline over the test region of 540−750 nm. This target testing provided also the $\langle \mathbf{V}_M|$ combination coefficients for M, constituting one vertex of the KLM triangle on the stoichiometric plane. The PCR value of 29.7% (as compared to the input 30%) was obtained from eq 30.

**Self-Modeling to Obtain the K and L Spectra.** The extrapolation of the multiexponential fit to zero time yielded the combination coefficients of K in the rank 5 decomposition of the whole data matrix. Multiplying the spectral eigenvectors (of the whole matrix) by these coefficients resulted in the output K spectrum. To use this in the rank 3 approximation of the first 19 difference spectra, the dot product of the 3 spectral eigenvectors (of the truncated matrix) with the K difference spectrum yielded $\langle \mathbf{V}_K|$, i.e., the K vertex of the KLM triangle.

The 19 data points, as represented by the 19 rows (3 elements each) of the **V** matrix are plotted in Figure 8. Visual inspection of these points or the corresponding difference spectra shows that point 18, for instance, contains all three intermediates. This point was used to remove the contribution of K and thereby finding the LM side of the stoichiometric triangle along the dashed line in Figure 8. Various test regions in the red were applied, and the best baseline for the resulting L + M mixture was obtained when this mixture was required to have no absorption above 690 nm. Moving along the LM side should remove all contribution by M and yield the L vertex. The criterion was to cancel the absorption of M in the blue region of 390−430 nm. The average absorption of L in this region was assumed to be the same as that of BR, in accordance with experience (and with the properties of the input spectra of the simulation, which were derived from experimental spectra). A certain range of deviation of the L absorbance from the BR absorbance in the blue range could be allowed, which would result in L spectra with slightly varying shape and amplitude.

Due to the input time constants of the simulation, the first 9 difference spectra contain no contribution by M, as also seen by the lining up of the corresponding data points along the KL side in Figure 8. This coincidence could be used to find the L vertex directly; however, the method described above should also be applicable when M contributes already to the earliest difference spectra.

**Self-Modeling to Obtain the N and O Spectra.** The K, L, and M spectra were used to fit the difference spectra to obtain the concentrations of these intermediates. From the trend of its decay, it is expected that the 24th difference spectrum contains an already negligible amount of K. Therfore, spectra 24−41 constitute a 4 component system of L, M, N, and O, and the L and M spectra are already known. The stoichiometric analysis has shown that up to the 33rd time point no recovery of BR is expected. It is possible to estimate the kinetics of BR recovery, as 1 minus the kinetics of the total intermediate concentration, by considering the facts that the latter should be a linear combination of the exponential functions found in the global multiexponential fit in such a way that this linear combination yields all ones up to the 33rd time point or, more conservatively, up to, say, time point 30. This linear combination was determined by a linear least-squares fit of the combination of the 8 exponentials to $|\mathbf{e}_{30}\rangle$. The amplitudes of the fastest 5 exponentials were negligible, so that the recovery kinetics of BR were found to consist of the three slow exponential components with amplitudes 0.066, −1.45, and 2.38. It is worth stressing that this time dependence is considerably steeper than a single exponential. Comparison with the input simulated BR kinetics over the entire time range was excellent. The possibility to estimate the kinetics of BR recovery is another clear advantage of EFASM compared to SM or other earlier analysis methods.

The difference spectra 31 through 41 were normalized by the kinetics of recovery so that in the resulting new **V** matrix all 41 points were supposed to fall on the stoichiometric surface. Self-modeling to obtain the N and O spectra was carried out using columns 24 through 41 of the normalized data matrix. SVD of this 18 column matrix yielded a rank of 3. The presence of the previously determined K, L, and M difference spectra was checked by target testing: The comparison of the difference spectrum $|\mathrm{K}\rangle$ with $\mathbf{U}(\mathbf{U}^{\mathrm{T}}|\mathrm{K}\rangle)$, and similar comparisons for L and M showed that the vectors $|\mathrm{K}\rangle$ and $|\mathrm{L}\rangle$ are not in the space spanned by the 18 column data matrix, whereas $|\mathrm{M}\rangle$ is. In other words, the contribution of the K and L intermediates at these late times is already negligible, in accordance with the input kinetics.

The determination of the MN line was performed in a similar fashion as the determination of the LM line in the search for the KLM triangle (see above). The contribution by O was removed while moving along the dashed line connecting the last (41st) data point with the 33rd data point (Figure 9, marked by arrow 1). The spectral range where no absorbance by either M or N was expected varied between 690−750 and 730−750

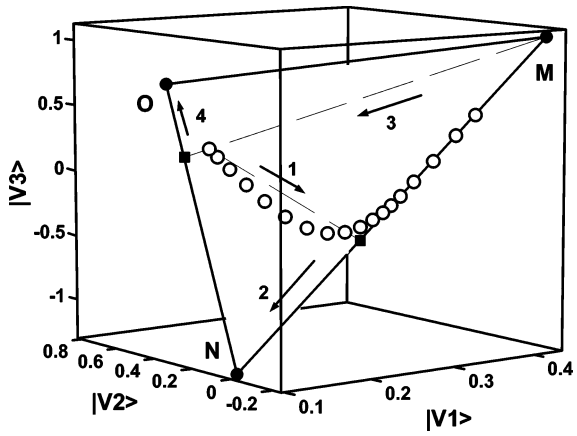


**Figure 9.** Self-modeling to obtain the N and O spectra. Open circles represent the rows of the **V** matrix from the rank 3 SVD of the last 18 columns of the data matrix, corrected for the BR recovery. The N vertex was located following arrow 1 and arrow 2, and the O vertex by following arrow 3 and arrow 4. For details see text.

nm, and the best range, with the smallest deviation of the resulting N spectrum from zero, turned out to be the 699−750 nm range. The N vertex was located along the MN line while monitoring the average absorbance of the resulting spectrum in the blue region (390−430 nm). A slightly higher average absorption of N than that of BR was allowed, in accordance with the input spectra.

To locate the NO line, another point is necessary. This could be obtained by removing all contribution due to M to the last difference spectrum when moving along the dashed line marked by arrow 3. A slightly lower absorption of O than that of BR was allowed in the blue region, again in accordance with the input spectra. The vertex of O is the most ambiguous one, because any point past the NO mixture (square in Figure 9) along the NO line could, in principle qualify. The obvious criterion of no negative absorption of O would yield an O spectrum which is too narrow and tall, unlike the commonly found spectra of retinal proteins. A safe guess is to expect a full width at half-maximum not too different from that of the BR spectrum. In the self-modeling, a 5% larger fwhm for O was assumed on the wavelength scale, which is again in accordance with the input spectra. A vertex too far from the 41st data point would result in a mixture of N and O in the last points of the kinetics with an overestimate of the amount of N. A vertex too close to the 41st data point, on the other hand, would underestimate the amount of N in the last data points. EFASM is unable to determine with 100 percent certainty the true spectrum of O and, therefore, the true kinetics of O and N, only a range of spectra and kinetics, which depends on the acceptable range of widths and heights of the O spectrum. The estimate of a 5% wider O spectrum (fwhm in wavelength) seems a reasonable guess based on our earlier studies, from which the input simulated data of this analysis were derived.

**Output Spectra and Kinetics.** The output spectra corresponding to the vertexes of the KLM and MNO triangles (Figures 8 and 9) are plotted in Figure 1, bottom. These spectra match the corresponding input spectra within error. Therefore, the output kinetics, which can be obtained by a linear least-squares fit of the data matrix (either the original, the SVD reconstructed or the exponential fit reconstructed) with the output difference spectra, agree also excellently with the input kinetics (not shown).

The output kinetics do not resolve the kinetics of the intermediate substates $L_1$ and $L_2$, $M_1$ and $M_2$, and $N_1$ and $N_2$, only their respective sums. The presence of substates may be conspicuous already from a visual inspection of the output kinetics (similar to the input kinetics in Figure 2, black lines), and the successful fit by the true photocycle scheme will faithfully reproduce the kinetics of the substates as well. The number of exponentials obtained from the multiexponential fit must be considered in determining the complexity of the model, which should contain the 5 spectrally distinct intermediates K, L, M, N, and O, whose spectra had been calculated in the process of EFASM and, in the case of this simulation, another 3 substates to account for the total number of intermediates. There are 35 different ways of selecting three additional substates from the 5 intermediates. If, however, the kinetics of the five spectrally distinct intermediates follow the trend shown in Figure 2 (black lines), one may suspect that L, M, and N may have substates, which would reduce the possibilities to 10. Although the most likely candidates would be schemes with two L, two M, and two N intermediates, it is difficult to rule out other possibilities from the visual inspection of the output intermediate kinetics. Further complication arises when one considers that there can be various pathways between the eight intermediates, other than the assumed linear reversible scheme of the simulation. If, in the case of real data, the global multiexponential fit yields 7−8 components, the selection of the true photocycle scheme with the correct number and distribution of intermediate substates and the true interconnections between them seems a rather unlikely coincidence. The determination of the intermediate spectra and their kinetics by EFASM helps considerably both by ensuring acceptable spectra and by guiding the modeling based on inspection of the kinetics of the intermediates.

## Discussion

Numerous articles have attempted to resolve the photocycle of bacteriorhodopsin from visible and/or vibrational kinetic spectroscopic data and find the true natural decomposition consisting of the extinctions and kinetics of the intermediates.[27,36−38,42,43] It appears that despite all these efforts there is still no consensus regarding the photocycle scheme. The majority of investigators accept a single sequence with reversible reactions and perhaps branches, but this model has also been repeatedly challenged, even as recently as in the year 2003.[16] There is also an ongoing debate regarding the best mathematical or computational method to dissect the available noisy experimental data.

Proponents of the various methods have claimed an unbiased analysis, but in my opinion, each protocol contains some ambiguity at a certain stage. SVD-EFASM, as described here, tries to eliminate this ambiguity as much as possible, as compared to previous works where the spectra were selected manually, over a grid search, or with the Monte Carlo method.[5,22] The roots of this inherent ambiguity lie in the severe spectral and kinetic overlap of the intermediates: the vertexes, sides, or limiting surfaces of the stoichiometric polyhedron fail to realize experimentally, and it is impossible to formulate exact criteria in the search for the pure intermediate spectra.

A global multiexponential fit alone assumes all unidirectional reactions, otherwise the resulting spectra and rates have no physical meaning. The direct photocycle model fits to 3D data (for example wavelength, time, and temperature[19]) are based on the determination of the number of intermediates from the number of exponentials found and rely on the assumption of temperature independent spectra. The rank of the data matrix, which may or may not be equal to the number of intermediates,

can be estimated from SVD. In his thorough review, Dioumaev[30] argues that SVD is a rank reducing procedure and is therefore bound to underestimate the number of photocycle intermediates from noisy data, which is correctly determined from a global multiexponential fit. He also argues that there is a direct conflict between retaining $r$ components from the SVD output and then fitting more than $r$ exponentials to the **V** matrix. From my analysis, it appears that this conflict can be resolved if one assumes, or accepts, the possibility of rank deficiency of the third type (see above), when matrix **c** is full rank but matrix $\epsilon$ is not. In case of rank deficiency of the second type, the very rapid rates defining the equilibrium between intermediates may manifest in a phenomenological rate constant well beyond the time resolution of the experiment, and therefore remain undetected, rendering the models deduced from the multiexponential fit deficient. This scenario will be detected by the impossibility of finding pure intermediate spectra without cross-contamination in the process of model fitting, no matter how complex the model may be.

The requirement of acceptable intermediate *absolute* spectra has been neglected in some cases, whereas in my opinion, this should be the cornerstone in the evaluation of the success of the analysis. SVD-EFASM is capable of demonstrating whether spectra of the required attributes can be deduced or, due probably to rank deficiency, only spectra of mixtures can be isolated. Potential spectra obtained by self-modeling yield kinetics which are expected to approximate the true kinetics the better the spectra are. These kinetics could guide any further efforts to resolve the photocycle by fitting of reaction schemes of the required complexity.

**Acknowledgment.** The author is indebted to Drs. András Dér and László Fábián for a critical reading of the manuscript. This work was supported by the Hungarian Scientific Research Fund (OTKA T034745).

## References and Notes

(1) Dedicated to the memory of my father, Dr. László Zimányi, Sr.
(2) Lanyi, J. K. *J. Phys. Chem. B* **2000**, *104*, 11441−11448.
(3) Heberle, J. *Biochim. Biophys. Acta* **2000**, *1458*, 135−147.
(4) Atkinson, G. H.; Ujj, L.; Zhou, Y. D. *J. Phys. Chem. A* **2000**, *104*, 4130−4139.
(5) Gergely, C.; Zimányi, L.; Váró, G. *J. Phys. Chem. B* **1997**, *101*, 9390−9395.
(6) Maeda, A. *Isr. J. Chem.* **1995**, *35*, 387−400.
(7) Dioumaev, A. K. *Biochemistry (Moscow)* **2001**, *66*, 1570−1579.
(8) Pebay-Peyroula, E.; Neutze, R.; Landau, E. M. *Biochim. Biophys. Acta* **2000**, *1460*, 119−132.
(9) Luecke, H. *Biochim. Biophys. Acta* **2000**, *1460*, 133−156.
(10) Lanyi, J. K.; Schobert, B. *J. Mol. Biol.* **2003**, *328*, 439−450.
(11) Brown, L. S. *Biochim. Biophys. Acta* **2000**, *1460*, 49−59.
(12) Stavenga, D. G.; Smits, R. P.; Hoenders, B. J. *Vision Res.* **1993**, *33* (8), 1011−1017.
(13) Hendler, R. W.; Dancsházy, Zs.; Bose, S.; Shrager, R. I.; Tokaji, Zs. *Biochemistry* **1994**, *33*, 4604−4610.
(14) Tokaji, Zs. *FEBS Lett.* **1998**, *423*, 343−346.
(15) Zimányi, L.; Váró, G.; Chang, M.; Ni, B.; Needleman, R.; Lanyi, J. K. *Biochemistry* **1992**, *31*, 8535−8543.
(16) Shrager, R. I.; Hendler, R. W. *J. Phys. Chem. B* **2003**, *107* (7), 1708−1713.
(17) Xie, A. H.; Nagle, J. F.; Lozier, R. H. *Biophys. J.* **1987**, *51*, 627−635.
(18) Groma, G. I.; Bogomolni, R. A.; Stoeckenius, W. *Biochim. Biophys. Acta* **1997**, *1319*, 59−68.
(19) Nagle, J. *Biophys. J.* **1991**, *59*, 476−487.
(20) Szundi, I.; Lewis, J. W.; Kliger, D. S. *Biophys. J.* **1997**, *73*, 688−702.
(21) Szundi, I.; Swartz, T. E.; Bogomolni, R. A. *Biophys. J.* **2001**, *80*, 469−479.
(22) Zimányi, L.; Lanyi, J. K. *Biophys. J.* **1993**, *64*, 240−251.
(23) Zimányi, L.; Kulcsár, Á.; Lanyi, J.; Sears, D.; Saltiel, J. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4408−4413.
(24) Zimányi, L.; Kulcsár, Á.; Lanyi, J.; Sears, D.; Saltiel, J. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4414−4419.
(25) Kulcsár, Á.; Saltiel, J.; Zimányi, L. *J. Am. Chem. Soc.* **2001**, *123*, 3332−3340.
(26) Zimányi, L. *Biopolym. (Biospectrosc.)* **2002**, *67*, 263−266.
(27) Hessling, B.; Souvignier, G.; Gerwert, K. *Biophys. J.* **1993**, *65*, 1929−1941.
(28) Aartsma, T. J.; Gouterman, M.; Jochum, C.; Kwiram, A. L.; Pepich, B. V.; Williams, L. D. *J. Am. Chem. Soc.* **1982**, *104*, 6278−6283.
(29) Nagle, J. F.; Zimányi, L.; Lanyi, J. K. *Biophys. J.* **1995**, *68*, 1490−1499.
(30) Dioumaev, A. K. *Biophys. Chem.* **1997**, *67*, 1−25.
(31) Onsager, L. *Phys. Rev.* **1931**, *37*, 405−426.
(32) Groma, G. I.; Dancsházy, Zs. *Biophys. J.* **1986**, *50*, 357−366
(33) Henry, E. R.; Hofrichter, J. *Methods Enzymol.* **1992**, *210*, 129−192.
(34) Hendler, R. W.; Shrager, R. I. *J. Biochem. Biophys. Methods* **1994**, *28*, 1−33.
(35) Dancsházy, Zs.; Tokaji, Zs. *Biophys. J.* **1993**, *65*, 823−831.
(36) Chizhov, I.; Chernavskii, D. S.; Engelhard, M.; Mueller, K.-H.; Zubov, B. V.; Hess, B. *Biophys. J.* **1996**, *71*, 2329−2345.
(37) Hendler, R. W.; Shrager, R. I.; Bose, S. *J. Phys. Chem. B.* **2001**, *105*, 3319−3328.
(38) Váró, G.; Lanyi, J. K. *Biochemistry* **1991**, *30*, 5016−5022.
(39) Zimányi, L.; Cao, Y.; Needleman, R.; Ottolenghi, M.; Lanyi, J. K. *Biochemistry* **1993**, *32*, 7669−7678.
(40) Ames, J. B.; Mathies, R. A. *Biochemistry* **1990**, *29*, 7181−7190.
(41) Gergely, C.; Ganea, C.; Groma, G.; Váró, G. *Biophys. J.* **1993**, *65*, 2478−2483.
(42) van Stokkum, I. H. M.; Lozier, R. H. *J. Phys. Chem. B* **2002**, *106*, 3477−3485.
(43) Borucki, B.; Otto, H.; Heyn, M. P. *J. Phys. Chem. B* **1999**, *103*, 6371−6383.