Automatic Factorization of Biological Signals by using Boltzmann Non-negative Matrix Factorization

Kenji Watanabe, Akinori Hidaka, and Takio Kurita

Abstract— **We propose an automatic factorization method for time series signals that follow Boltzmann distribution. Generally time series signals are fitted by using a model function for each sample. To analyze many samples automatically, we have to apply a factorization method. When the energy dynamics are measured in thermal equilibrium, the energy distribution can be modeled by Boltzmann distribution law. The measured signals are factorized as the non-negative sum of the probability density function of Boltzmann distribution. If these signals are composed from several components, then they can be decomposed by using the idea of non-negative matrix factorization (NMF). In this paper, we modify the original NMF to introduce the probability density function modeled by Boltzmann distribution. Also the number of components in samples is estimated by using model selection method. We applied our proposed method to actual data that was measured by fluorescence correlation spectroscopy (FCS). The experimental results show that our method can automatically factorize the signals into the correct components.**

I. INTRODUCTION

Factorization of time series signals is very important in biological researches, such as spike analysis in brain science [1] and analysis of the protein dynamics in molecular biology [2], [3]. Especially, in the field of molecular biology, Fluorescence Correlation Spectroscopy (FCS) [4]-[6] begins to be often used to measure and analyze the protein dynamics in living cell [2], [3]. Such analysis of time series signals would be more important in the future.

In the case of FCS, the time series signals are represented as a linear combination of positive components constructing from the time constant and the existence ratio when the energy dynamics in thermal equilibrium of the target is measured. In the previous analysis methods, these signals are fitted as a linear combination of the model functions. The analytical results such as the time constants are estimated from each signal sample. These methods are excellent to analyze a specific phenomenon in one signal sample because these model functions are constructed for each target. However, it is often required to get statistical tendencies of many samples rather than single sample. In the analysis of the statistical tendency by using the previous methods, the validities of fitting results and the necessity of another statistical analysis are manually checked. These manual checks probably make the analysis results arbitrary. We can automate the analysis of time series signals by an automatic factorization. The statistical tendency of many samples can be estimated by taking the average of the decomposed components. In the analysis of actual measurement data, it is very important to reduce the risk of the arbitrary decisions. But there are not so many the automatic analysis methods.

Automatic signal factorization, for example, factor analysis, independent component analysis (ICA) [7], [8], non-negative matrix factorization (NMF) [9], [10] has been examined in many fields. NMF was applied for the time series signals that composed from a linear combination of the non-negative existence ratios and time constants [11]. NMF is useful for factorization of the time series signals because NMF is guaranteed the non-negativity of outputs in case of the non-negative input signals. However, NMF has some inconveniences such as the factorization results depend on the initial values of factorization process, and the decomposed factors are not guaranteed to follow the given model.

In this paper, we modified NMF to introduce a model function of the basis vectors. The model function is assumed as the probability density function that represents Boltzmann distribution in order to represent the physical phenomenon of the signals. Our modified NMF is named Boltzmann Non-negative Matrix Factorization (BzNMF). The number of components is also estimated by using cross validation (CV) [12]. CV is one of the well-known model selection methods and it is able to select the number of components.

To verify the validity of our proposed method, we applied the proposed method to the actual measurement data obtained by using fluorescence correlation spectroscopy (FCS) [4]-[6]. Recently FCS becomes gradually popular to measure and analyze the protein dynamics in living cell [2], [3]. In the analysis of signals such as FCS measurement data, an estimation method [13], [14] has been developed in the molecular biological field. This data is suitable for our BzNMF.

Kenji Watanabe is with the Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba-shi, Ibaraki-ken, 305-8577 Japan (phone: 029-861-5080; e-mail: kenji-watanabe@aist.go.jp).

Akinori Hidaka is with the Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba-shi, Ibaraki-ken, 305-8577 Japan (e-mail: hidaka.akinori@aist.go.jp).

Takio Kurita is with the National Institute of Advanced Industrial Science and Technology (AIST), AIST Central 2, 1-1-1 Umezono, Tsukuba-shi, Ibaraki-ken, 305-8568 Japan (e-mail: takio-kurita@aist.go.jp).

978-1-4244-1821-3/08/$25.00©2008 IEEE

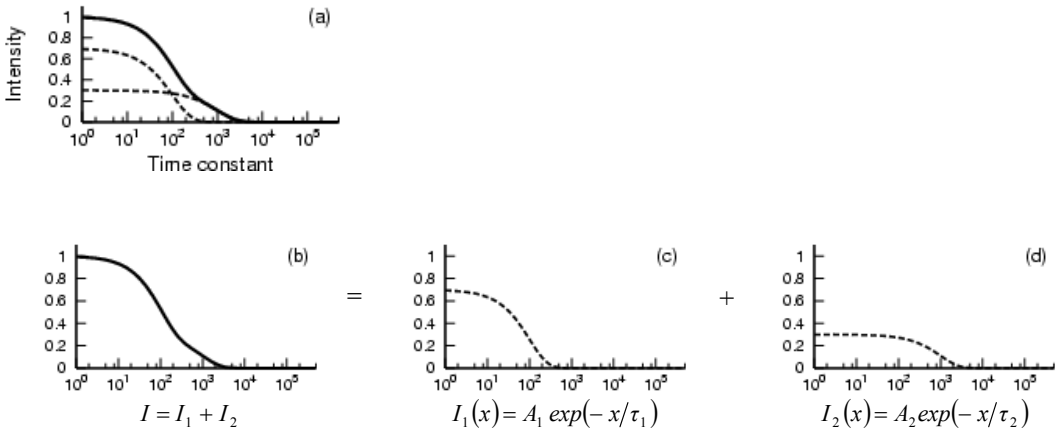

Fig. 1. Factorization of the energy distribution. Each component of the energy distribution is modeled by Boltzmann distribution law (a). The measurement energy distribution (b) can be represented as the sum of each component's energy distributions (c, d). Lines show the total measurement energy distribution (a, b). Broken lines show the each component that is represented the probability density function of Boltzmann distribution (a, c, d).

II. ANALYSIS OF ENERGY DISTRIBUTION

Generally, the energy distribution of time series signals in thermal equilibrium such as a motion of particles or an energy migration can be modeled by Boltzmann distribution law. Autocorrelation function (ACF) calculated from these time series signals is similar to the energy distribution. Thus we can assume that ACF is also represented as Boltzmann distribution.

The values of ACF calculated from a measurement sample can be represented as a vector format (called a feature vector). A set of feature vectors is obtained from many samples measured with the equal condition and the set of vectors constructs an ACF matrix.

Fig. 1 shows the concept of the energy distribution. Here I is the energy intensity and I_i is the energy intensity of the i'th component. If ACF is calculated from the time series signals that follow Boltzmann distribution, the values of ACF show the energy intensity as shown in Fig. 1 (a, b) with lines. Generally the total energy intensity can be represented as a linear combination of the component's energy intensities as shown in Fig. 1 (b, c, d). The component's intensities also follow Boltzmann distribution. They can be represented by using the probability density function as shown in Fig. 1 (c, d). This probability density function is modeled by the equation shown in Fig. 1 (c, d). Here A_i is the amplitude and τ_i is the time constant of the i'th component respectively.

To analyze the dynamics of such samples, the fitting of a model function to each vector is often used. However, we often have to analyze the statistical tendency of the dynamics in many samples. Factorization methods are able to decompose the data matrix into the components (the basis vectors) that reflect the behaviors of all samples.

Since the ACF matrix represents the physical phenomenon, we have to decompose the matrix with the non-negative components. This means that the input vector of ACF matrix must be represented as a linear combination of these components. Also each of the estimated components must be modeled by the probability density function of Boltzmann distribution. NMF [9], [10] was proposed to decompose a given non-negative matrix into a non-negative basis matrix and a coefficient matrix. But the output of NMF is not guaranteed to agree with the given model function. Therefore we introduce the model function to the basis vectors of NMF. As a result, the outputs of BzNMF are represented as the model function.

III. PREVIOUS METHOD

A. Fluorescence Correlation Spectroscopy

We use FSC as the actual measurement data to verify our proposed method. FCS is one of the techniques to measure the fluorescence intensity fluctuations caused by fluorescent probe movement of free diffusion. FCS measurement data is to deduce diffusion times and existence ratios of fluorescent probes from ACF that was calculated from the fluorescence intensity fluctuations. ACF is defined as follows:

$$v(x) = \frac{< I_t I_{t+x} >}{< I >^2}, \qquad (1)$$

where I_t is the fluorescence intensity in time t. Diffusion time x is defined as $x = \Delta t$. $< I >^2$ is the square of the averaged fluorescence intensity.

Since ACF may include several positive components related with different origins, usually the obtained ACFs are fitted by one-, two-, or three-components model as follows:

$$v(x) = 1 + \frac{1}{N}\sum_i F_i \left(1 + \frac{x}{\tau_i}\right)^{-1} \left(1 + \frac{x}{s^2 \tau_i}\right)^{-1/2}, \qquad (2)$$

where F_i and τ_i are the fraction and diffusion time of component i, respectively. N is the number of fluorescence molecules in the detection volume element defined by $s = z_0 / w_0$, radius w_0 and length $2z_0$. The correlation amplitude of the function (y intercept, the value of $v(0)$) is determined by the reciprocal of the number of fluorescence molecules in detection volume. For example, in [2] ACF of rhodamine 6G (Rh6G) water solution were measured for 30s five times at 10s interval, then the diffusion time (τ_{Rh6G}) and s were obtained by one-component fitting of the measured ACF in each sample.

IV. NON-NEGATIVE MATRIX FACTORIZATION AND BOLTZMANN NON-NEGATIVE MATRIX FACTORIZATION

A. Non-negative Matrix Factorization

We have to decompose the matrix with the non-negative coefficients when we want to decompose the ACF matrix into the components (the basis vectors) because both ACF and the basis vectors are non-negative. NMF [9], [10] was proposed to decompose a given input matrix $\boldsymbol{V} \in \Re^{n \times m}$ into a basis matrix $\boldsymbol{W} \in \Re^{n \times r}$ and a coefficient matrix $\boldsymbol{H} \in \Re^{r \times m}$, as follows:

$$\boldsymbol{V} \approx \boldsymbol{WH}. \qquad (3)$$

This means that $\boldsymbol{WH}$ is an approximation of the input matrix $\boldsymbol{V}$.

In NMF the objective function is defined by the "divergence" of $\boldsymbol{V}$ from $\boldsymbol{WH}$ as the measure of cost for factorization. Here we use the objective function which minimizes the generalized Kullback-Leibler divergence. The objective function was given as follows:

$$D(\boldsymbol{V} \,||\, \boldsymbol{WH}) = \sum_{ij}\left(\boldsymbol{V}_{ij}\, log \frac{\boldsymbol{V}_{ij}}{(\boldsymbol{WH})_{ij}} - \boldsymbol{V}_{ij} + (\boldsymbol{WH})_{ij}\right). \qquad (4)$$

From (4), the multiplicative update rules of the basis and coefficient matrices in NMF were derived as follows:

$$\begin{cases} \boldsymbol{W}_{ia} \leftarrow \boldsymbol{W}_{ia} \sum_{\mu} \dfrac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}} \boldsymbol{H}_{a\mu} \\ \boldsymbol{W}_{ia} \leftarrow \dfrac{\boldsymbol{W}_{ia}}{\sum_j \boldsymbol{W}_{ja}} \\ \boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \sum_i \boldsymbol{W}_{ia} \dfrac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}} \end{cases}, \qquad (5)$$

where $\boldsymbol{V}_{i\mu}$ is the component of input matrix that is the i'th row of the μ'th sample, $\boldsymbol{W}_{ia}$ is the component of basis matrix that is the i'th row of the a'th rank and $\boldsymbol{H}_{a\mu}$ is the component of coefficient matrix that is the a'th rank of the μ'th sample. $(\boldsymbol{WH})_{i\mu}$ is the component of approximation matrix, i.e., $(\boldsymbol{WH})_{i\mu} = \sum_a \boldsymbol{W}_{ia} \boldsymbol{H}_{a\mu}$.

B. Boltzmann Non-negative Matrix Factorization

There is no guarantee to reflect a physical phenomenon in the basis matrix of original NMF. We propose to introduce the constraint of the probability density function of Boltzmann distribution into the basis vectors. The constraints of this physical phenomenon is defined as $\boldsymbol{W}_{ik} = \boldsymbol{a}_k \, exp(-\boldsymbol{x}_i / \boldsymbol{\tau}_k)$. The functional form of this constraint is derived from the probability density function of Boltzmann distribution. Also there is another constraint of the total sum as $\sum_i \boldsymbol{W}_{ik} = 1$. The approximation matrix and the objective function of BzNMF are the same as the original NMF shown in (3) and (4) respectively. From the objective function, by taking the differential with respect to $\boldsymbol{\tau}_k$ becomes:

$$\frac{\partial D}{\partial \boldsymbol{\tau}_k} = -\frac{\boldsymbol{a}_k}{\boldsymbol{\tau}_k^{\,2}} \sum_{ij} \frac{\boldsymbol{V}_{ij}}{(\boldsymbol{WH})_{ij}} \boldsymbol{x}_i \, exp\left(-\frac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right) \boldsymbol{H}_{kj} + \frac{\boldsymbol{a}_k}{\boldsymbol{\tau}_k^{\,2}} \sum_{ij} \boldsymbol{x}_i \, exp\left(-\frac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right) \boldsymbol{H}_{kj}. \qquad (6)$$

We can derive the multiplicative update rule of $\boldsymbol{\tau}_k$ by using the constant $\eta(\boldsymbol{\tau}_k)$ of the gradient descent formula $\boldsymbol{\tau}_k \leftarrow \boldsymbol{\tau}_k - \eta(\boldsymbol{\tau}_k)\{\partial D / \partial \boldsymbol{\tau}_k\}$. The constraint $\eta(\boldsymbol{\tau}_k)$ is as follows:

$$\eta(\boldsymbol{\tau}_k) = \frac{\boldsymbol{\tau}_k}{\frac{\boldsymbol{a}_k}{\boldsymbol{\tau}_k^{\,2}} \sum_{ij} \boldsymbol{x}_i \, exp\left(-\frac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right) \boldsymbol{H}_{kj}}. \qquad (7)$$

This gives the update rule for $\boldsymbol{\tau}_k$. Similarly we can derive the multiplicative update rules for $\boldsymbol{a}_k$. Thus the multiplicative update rules in BzNMF are given as follows:

$$\begin{cases} \boldsymbol{H}_{k\mu} \leftarrow \boldsymbol{H}_{k\mu} \sum_i \dfrac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}} \boldsymbol{a}_k \exp\left(-\dfrac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right) \\ \boldsymbol{\tau}_k \leftarrow \boldsymbol{\tau}_k \dfrac{\sum_{ij} \dfrac{\boldsymbol{V}_{ij}}{(\boldsymbol{WH})_{ij}} \boldsymbol{x}_i \exp\left(-\dfrac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right) \boldsymbol{H}_{kj}}{\sum_{ij} \boldsymbol{x}_i \exp\left(-\dfrac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right) \boldsymbol{H}_{kj}} \\ \boldsymbol{a}_k \leftarrow \dfrac{1}{\sum_i \exp\left(-\dfrac{\boldsymbol{x}_i}{\boldsymbol{\tau}_k}\right)} \end{cases}, \quad (8)$$

where $\boldsymbol{\tau}_k$ and $\boldsymbol{a}_k$ are the time constant and amplitude of rank k, respectively. $\boldsymbol{x}_i$ is the i'th time of the time series. The time constant and amplitude are non-negative. The existence ratio of the k'th rank in the μ'th sample is represented as $\boldsymbol{H}_{k\mu}\boldsymbol{a}_k \big/ \sum_l \boldsymbol{H}_{l\mu}\boldsymbol{a}_l$. The component of the k'th rank in the μ'th sample is represented as $\boldsymbol{W}_{ik}\boldsymbol{H}_{k\mu}$. The proof of these multiplicative update rules are the same as original NMF [10].

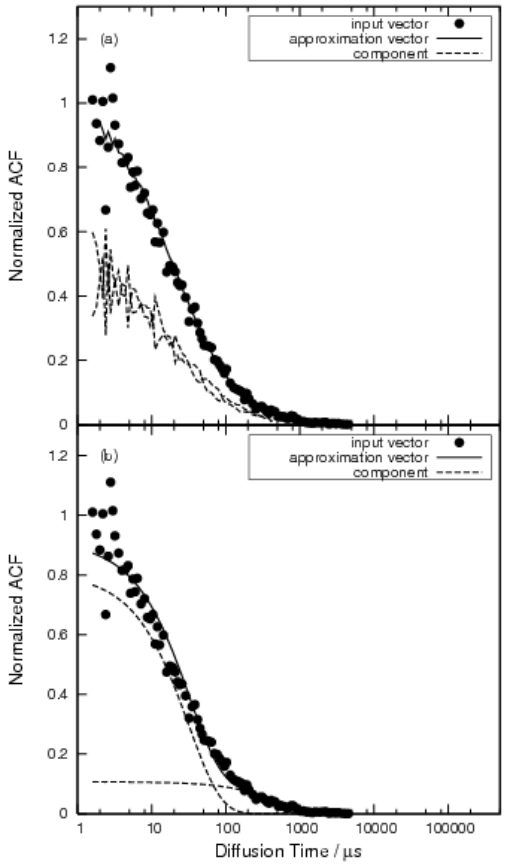

Fig. 2. Comparison of factorization result. The samples of Rh6G were factorized by using NMF (a) and BzNMF (b). Closed circles show the input vector that represents the ACF of Rh6G (a, b). Lines show the results of approximation vector by using NMF and BzNMF (a, b). Broken lines show the components that represent the products of basis vector and coefficient in each rank (a, b).

V. RANK ESTIMATION

A. Cross Validation

The number of basis vectors (components), namely the rank, is usually manually decided. But automatic determination of the rank is important because it has influence to the results of decomposition. In this paper, we use *K*-fold CV [12] to decide the rank. CV is one of the popular techniques of model selection. It evaluates models more directly than other theoretical methods such as AIC, MDL and etc. It can be used to solve the problem of over-fitting. The technique of *K* -fold CV can be summarized as follows

1) The training samples are arbitrarily divided into *K* subsets.

2) The one of subsets is left out for evaluation and the model is fitted using the samples in the remaining *K*-1 subsets. The mean square errors of the subset that left out for evaluation is calculated using the trained model. Since there are *K* possibilities for how to leave out the subset for evaluation, the mean square errors of all *K* subsets are evaluated.

3) The final generalization performance is computed as the average of the mean square errors for the *K* subsets.

We can define *K*-fold CV value as follows:

$$CV = \frac{1}{(m/K)} \sum_{i=1}^{(m/K)} CV(i), \quad (9)$$

$$CV(i) = \frac{1}{L} \sum_{j=1}^{L} \left\| \boldsymbol{V}_i \log \frac{\boldsymbol{V}_i}{(\boldsymbol{WH})_j} - \boldsymbol{V}_i + (\boldsymbol{WH})_j \right\|, \quad (10)$$

where $\left\| \boldsymbol{V}_i \log\left(\boldsymbol{V}_i / (\boldsymbol{WH})_j\right) - \boldsymbol{V}_i + (\boldsymbol{WH})_j \right\|$ is the L1 norm and $CV(i)$ is the CV value of i'th input vector $\boldsymbol{V}_i$. $\boldsymbol{V}_i$ is the i'th sample of the (m/K) samples. (m/K) is a real number. $(\boldsymbol{WH})_j$ is the j'th approximation vector that calculated from the L samples. We define that $L = (m\text{-}m/K)$. Then the number of basis vectors is estimated from the model with the minimum CV value. Corresponding with the objective function of BzNMF (4), we define the measure on CV evaluation by using the generalized Kullback-Leibler divergence. In the following experiments, we use 10-fold CV.

VI. EXPERIMANTAL RESULTS

A. Decomposition of Basis vectors

NMF is not always possible to factorize the time series signals because NMF has no constraint on the model function that represents physical phenomena. But in the proposed BzNMF, the constraint on Boltzmann distribution is

introduced. The factorization results by using the original NMF and BzNMF are shown in Fig. 2 so as to compare the factorization results.

The sample was rhodamine 6G (Rh6G) water solution measured by FCS and the concentration of Rh6G was 10^{-7}mol/l. It had 54 samples and the sample was 92 dimensional vector ($1.6 \le \boldsymbol{x}_i \le 4505.6$). Rh6G is the chemical particle that exists as the one-component in water. We can assume that the Rh6G exists as the ideal state when the concentration of Rh6G water solution is sufficiently low level. The sample is represented as ACF and was calculated from the time series signal by using (1). These ACFs of Rh6G were normalized by using linear regression method. In this experiment the rank was set to 2 because the signals probably contain the main component and some artifacts.

It is noticed that the basis vectors obtained by the original NMF are not smooth and two components are similar as shown in Fig. 2 (a). On the other hand, BzNMF could clearly decompose to the basis vectors that reflect the model function as shown in Fig. 2 (b). The averaged existence ratio and diffusion time of the main component calculated by BzNMF were 0.8840 and 31.83μs respectively. This diffusion time of main component is sufficiently accurate because the averaged diffusion time by using (2) was estimated 24.89 ±11.49μs. Therefore, we can say that in comparison with the original NMF, BzNMF can clearly factorize the samples that follow Boltzmann distribution. Also the factorization results of BzNMF are similar to the results obtained by the manual fitting by using (2).

B. Rank estimation

If the rank of samples has been known, BzNMF can accurately estimate the parameters as in the case of Fig. 2. But in the analysis of actual measurement data, the rank is often unknown. Hence, the rank was estimated by using CV from the factorization results of BzNMF.

The results of rank estimation for Rh6G by using 10-fold CV are shown in Fig. 3. The samples were same as the section VI.*A*. From Fig. 3, we can estimate the rank as 5 because the CV values has minimum when the rank is equal to 5.

The factorization results for Rh6G samples with rank 5 are shown in Fig. 4. The samples were same as the section VI.*A*. From Fig. 4, it is noticed that the approximation vector is more appropriate than the rank 2 factorization result in Fig. 2 (b). The 2nd component of the fast diffusion times was the main component because the existence ratio of 2nd component gave the highest value in this sample.

To make the components clear, the factorization results by

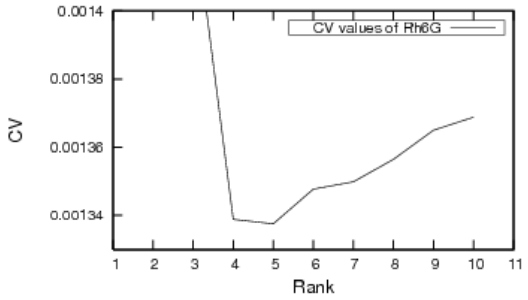

Fig. 3. Rank estimation by using 10-fold CV. Rank estimation result of Rh6G. Line shows the CV values in each rank.

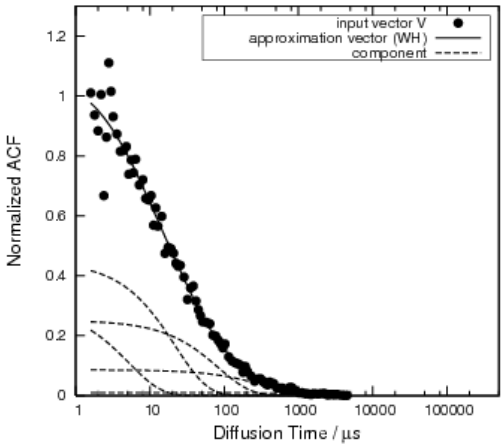

Fig. 4. Factorization result in rank 5. The samples of Rh6G were factorized by using BzNMF. Closed circles show the input vector that represents the ACF of Rh6G. Lines show the results of approximation vector by using BzNMF. Broken lines show the components that represent the products of basis vector and coefficient in each rank.

TABLE I
ANALYSIS RESULT OF RH6G

FCS analysis method [*1]		BzNMF	
Diffusion Time	Existence Ratio	Diffusion Time	Existence Ratio
7.23 ±13.95	(0.3283 ±0.2848)	5.21	(0.1883 ±0.0508)
24.89 ±11.49	(1.0)	22.07	(0.4800 ±0.0489)
		84.17	(0.2458 ±0.0157)
		360.67	(0.0753 ±0.0070)
		2472.56	(0.0108 ±0.0021)

[*1] The fitting of the model function by using (2). The 1st component represents the averaged diffusion time and existence ratio of the triplet state from all samples.

BzNMF and the fitting results of the current FCS analysis method are shown in Table I. The fitting of model function in the FCS analysis method was done by using (2) in each sample. This fitting method was fitted to the one-component (rank 1) because Rh6G exists as the one-component from the chemical knowledge. The fastest diffusion time obtained by the FCS analysis method represents the triplet state of Rh6G. The diffusion time and existence ratio by the FCS analysis method were averaged from all samples. The existence ratios of BzNMF were averaged from all samples.

In Table I, the 1st component of BzNMF is the similar with the result of the FCS analysis method's triplet state. The main component was obtained as the 2nd component in BzNMF. The main component probably represents the diffusion of Rh6G in water solution. The other components slower than these two components may represent the artifacts, for example, a spectroscopy problem such as the measurement noise.

These results suggest that the proposed method can automatically factorize the signals that represent Boltzmann distribution. But it has the tendency in which the rank is larger than the FCS analysis method. Also the existence ratio of 3rd component is too large to neglect. The result of 3rd component by using BzNMF is probably affected to the slow diffusion samples when the large standard deviation of the main component's diffusion time by using the FCS analysis method is considered. We have to pay attention to these tendencies when the proposed method is applied to actual measurement data.

C. Application to the biological sample

We applied the proposed method to the actual measurement data of biological samples. The factorization of protein dynamics signal is very important to decrease the influence of arbitrary decision in molecular biology. The factorization of protein dynamics signal is more difficult than the chemical particle's. The automatic factorization of these signals is novel approach in the molecular biology. The actual measurement data by using FCS is Enhanced Green Fluorescence Protein (EGFP) in living cell. The data of EGFP were 45 samples. The analysis sample of EGFP was 120 dimensional vector ($2.2 \le \boldsymbol{x}_i \le 65536.0$) and these ACFs of EGFP were normalized by using linear regression method.

The analysis results of this FCS data by using the proposed method and the FCS analysis method are shown in Fig. 5. The fitting was performed using (2) in the FCS analysis method. The results of the FCS analysis method were fitted to the one-component (rank 1) because EGFP exists as the one-component and has hardly interaction with intracellular structures from the biological knowledge. In the results of the proposed method, the EGFP samples were factorized by using BzNMF and the rank was decided as 5 by using 10-fold CV. From Fig. 5, it is noticed that the 2nd component of the proposed method is the main component.

To compare the result of the FCS analysis method with the result of the proposed method, the diffusion times and the existence ratios are shown in Table II. The result of the FCS analysis method with the fastest diffusion time represented the triplet state of EGFP. The diffusion time and existence ratio of the FCS analysis method were averaged from all samples. The existence ratios of BzNMF were averaged from all samples. In Table II, the 1st component obtained by BzNMF is similar

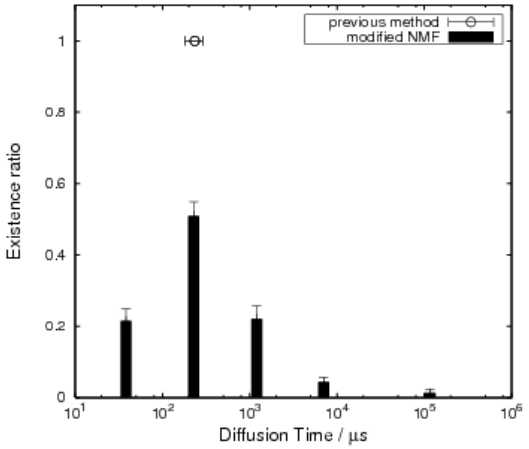

Fig. 5. Analysis results of EGFP. Open circles show the analysis results by using previous method. Bars show the analysis results by using BzNMF. Error bars show the standard deviations of the diffusion times of previous method or the existence ratios of BzNMF.

TABLE II
ANALYSIS RESULT OF EGFP

FCS analysis method *1		BzNMF	
Diffusion Time	Existence Ratio	Diffusion Time	Existence Ratio
32.70 ±26.73	(0.1687 ±0.0749)	38.27	(0.2144 ±0.0353)
243.57 ±53.44	(1.0)	228.01	(0.5090 ±0.0392)
		1198.01	(0.2205 ±0.0364)
		7015.59	(0.0438 ±0.0123)
		116154.57	(0.0125 ±0.0106)

*1 The fitting of the model function by using (2). The 1st component represents the averaged diffusion time and existence ratio of the triplet state from all samples.

with the result of the FCS analysis method's triplet state. The 2nd component is the main component because the 2nd component is similar to the result of the FCS analysis method. The main component probably represents the diffusion of EGFP in living cell. The other components with slower diffusion times may represent the artifacts, for example, a spectroscopy problem such as the measurement noise. But the 3rd component probably represents a weak interaction with intracellular structures because the diffusion time is 1198.01μs and the existence ratio is 0.2205 ±0.0364. The diffusion time of 3rd component is too slow if this component represents the free diffusion of the monomeric EGFP. Also the existence ratio of 3rd component represents the about half of the main component's existence ratio. From these reasons, it is considered that the 3rd component probably represents a weak interaction with intracellular structures.

These results suggest that the proposed method can automatically factorize the signals of protein dynamics and to make clear the tendency of components from many samples.

VII. CONCLUSION

Our proposed method is the automatic factorization method for the time series signals. In factorizing the signals that can be modeled as Boltzmann distribution, our BzNMF gives better decomposition results of the input matrix into the basis matrix than the original NMF. This is because we introduced the constraint on Boltzmann distribution into the NMF algorithm. We think that it is novel approach to introduce the constraint on physical phenomenon in the pattern recognition [2], [4] or the spectroscopy [12], [13]. Furthermore, our proposed method can estimate the rank by using 10-fold CV. The rank estimated by using CV has the tendency to give larger rank than the true rank for difficult samples. By using the rank estimation, the proposed method can factorize the unknown signals in the case of the signals represent the energy distribution.

In the analysis of model function fitting method, the results of the fitting are manually judged especially in molecular biology [2], [3], [14]. There is a risk that the analysis results of the model function fitting method probably reflect arbitrary decisions. To solve these problems, the automation of analysis is very important. In the factorization result of actual measurement data by using FCS, our method had the tendencies in which the rank is larger than the FCS analysis method and the factorized component is affected to the samples that has the slow diffusion component as shown in Table I. But our proposed method was able to estimate the sufficiently accurate results as the practical analysis method in comparison with the current FCS analysis method as shown in Fig. 5 and Table II. From these reasons, our proposed method is useful for the automatic factorization of the signals that follow Boltzmann distribution measured by FCS at least.

For future works, we will have to verify the results of time series signals by the proposed method quantitatively, and we want to propose the automatic recognition method about physical phenomena in the pattern recognition framework.

ACKNOWLEDGMENT

We are grateful to Prof. T. Miyazaki (Department of Bioresources, Hokkaido University Research Center for Zoonosis Control, Japan) and Associate Prof. M. Kinjo (Laboratory of Biophysics, Research Institute for Electronic Science, Hokkaido University, Japan) for the making of biological materials and the measurement of FCS.

REFERENCES

[1] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, Vol. 442, pp. 164-171, Jul. 2006

[2] K. Watanabe, K. Saito, M. Kinjo, T. Matsuda, M. Tamura, S. Kon, T. Miyazaki, and T. Uede, "Molecular dynamics of STAT3 on IL-6 signaling pathway in living cells," *Biochem. Biophys. Res. Commun.*, Vol. 324, pp. 1264-1273, 2004

[3] A. Kitamura, H. Kubota, C.-G. Pack, G. Matsumoto, S. Hirayama, Y. Takahashi, H. Kimura, M. Kinjo, R. I. Morimoto, and K. Nagata, "Cytosolic chaperonin prevents polyglutamine toxicity with altering the aggregation state," *Nature Cell. Biol.*, Vol. 8, No. 10, pp. 1163-1170, Oct. 2006

[4] M. Ehrenberg, and R. Rigler, "ROTATIONAL BROWNIAN MOTION AND FLUORESCENCE INTENSITY FLUCI'UATIONS," *Chem. Phys.*, 4 (1974) 390-401

[5] E. L. Elson, and D. Magde, "Fluorescence correlation spectroscopy. I. Conceptual basis and theory," *Biopolymers*, Vol. 13, pp. 1-27, Feb. 2004

[6] D. E. Koppel, "Statistical accuracy in fluorescence correlation spectroscopy," *Phys. Rev. A*, Vol. 10, No. 6, pp. 1938-1945, Dec. 1974

[7] P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994

[8] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *NeuroImage*, vol. 34, pp. 1443–1449, 2007

[9] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, Oct. 1999

[10] D. D. Lee, and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *Adv. Neural Info. Proc. Syst.*, vol. 13, pp. 556-562, 2001

[11] K. Watanabe, and T. Kurita, "Automatic Factorization of Biological Signals Measured by Fluorescence Correlation Spectroscopy using Non-negative Matrix Factorization," *ICONIP 2007 14th Int. Conf. on Neural Information Processing*, WMB-3, accepted.

[12] J. Shao, "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, Vol. 88, No. 422, pp. 486-494, Jun. 1993

[13] R. Rao, R. Langoju, M. Gösch, P. Rigler, A. Serov, and T. Lasser, "Stochastic Approach to Data Analysis in Fluorescence Correlation Spectroscopy," *J. Phys. Chem. A*, Vol. 110, No. 37, pp. 10674-10682, 2006

[14] H. D. Kim, G. U. Nienhaus, T. Ha, J. W. Orr, J. R. Williamson, and S. Chu, "Mg^{2+}-dependent conformational change of RNA studied by fluorescence correlation and FRET on immobilized single molecules," *Proc. Natl. Acad. Sci. USA*, Vol. 99, No. 7, pp. 4284-4289, Apr. 2002