LETTER

Automatic Factorization of Biological Signals Measured by Fluorescence Correlation Spectroscopy using Non-negative Matrix Factorization

Kenji Watanabe

Department of Computer Science, Graduate School of Systems and Informatoin Engineering,
University of Tsukuba,
1-1-1 Tennodai, Tsukuba-shi, Ibaraki-ken, 305-8577 Japan
E-mail: kenji-watanabe@aist.go.jp

Takio Kurita

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

(Submitted on December 12, 2007)

Abstract – This paper proposes an automatic factorization method of the biological signals measured by Fluorescence Correlation Spectroscopy (FCS). Since the signals are composed from several positive components, the signals are decomposed by using the idea of Non-negative matrix factorization (NMF). Each component is approximated by a model function and the signals are factorized as the non-negative sum of a few model functions. Analytical accuracy of the proposed method was verified by using biological data that were measured by FCS. The experimental results showed that the proposed method could automatically factorize the signals and could succeed to obtain the similar results with the manual investigations.

Keywords – Signal processing, NMF, Pattern recognition, Protein dynamics

1. 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], [5], [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 become more important in the future. However, the current FCS analysis method is not efficient because each sample is fitted as a linear combination of the model functions and the parameters of model functions are plotted to find the frequent components. In addition, the current analytical results of FCS have the possibility to include the arbitrary decision that means to reflect the subjectivity of researcher because the examination of analytical results and judgments of re-analysis are manually decided. To improve the current FCS analysis method, a model function [7] or an approximation method [8] has been modified. But these modifications were not sufficient because the researchers in this field want to know what components are included in the set of signals and what kind of the statistical tendency is found in the large amount of samples. In molecular biology, the components are manually found and its statistical tendency is investigated through statistical analysis. This process is time consuming and subjective.

Automatic signal factorization has been studied in a lot of fields, for example, factor analysis, principal component analysis (PCA), independent component analysis (ICA) [9], [10], non-negative matrix factorization (NMF) [11], [12]. Especially, NMF is probably most effective for the factorization of non-negative energy distribution such as a molecular dynamics in thermal equilibrium because this energy distribution can be

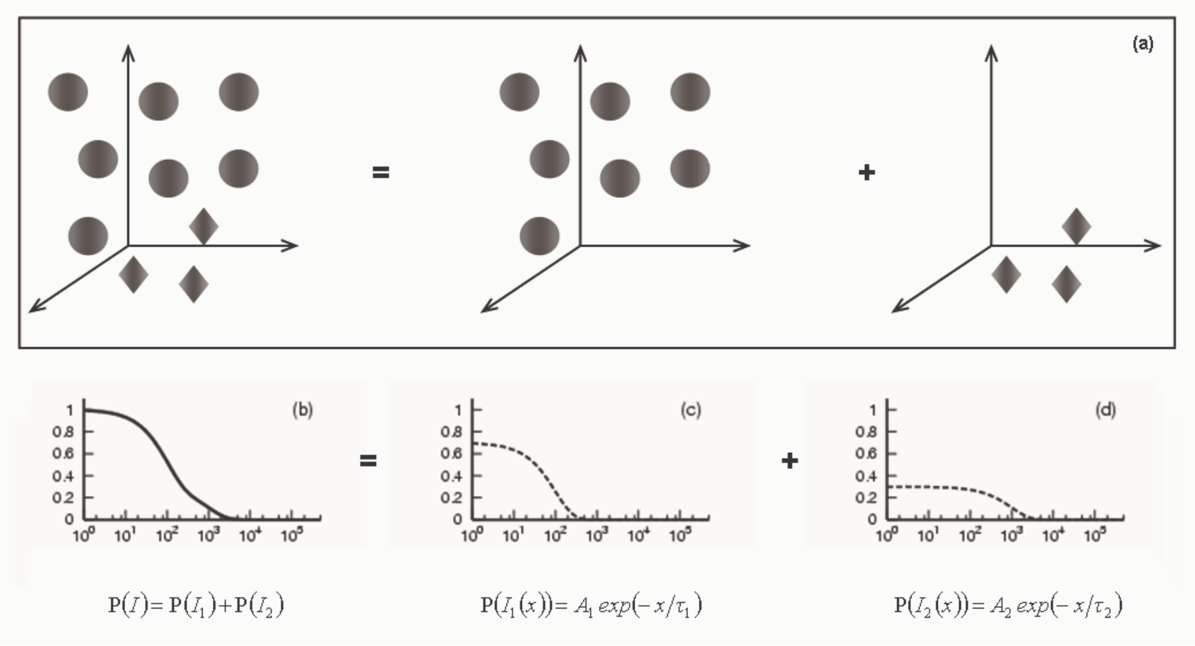

Figure 1. Schematic graph of the energy distribution of several particles. The energy distribution is modeled by Boltzmann distribution law. The schematic graph of low density system with the different particles is shown in (a). Circles represent the particle 1. Diamonds represent the particle 2. The energy distribution constructed from two components is shown with line in (b). The energy distributions of the particle 1 and particle 2 are shown with broken lines in (c) and (d), respectively. They are modeled by the probability density function of Boltzmann distribution.

represented by the non-negative sum of a few non-negative components. Also the components are not guaranteed to be orthogonal. On the other hand, PCA and ICA are not suitable for this application because they do not have non-negativity and decompose the signals into a sum of orthogonal basis vectors.

In this paper, we propose a factorization method of biological signals measured by FCS in which the idea of NMF is used to decompose the signals into several positive components. Each component is represented by a model function derived by considering its physical phenomena and is approximated through the nonlinear least squares method. By using NMF approach, we can directly find the components included in a set of signals obtained from the many samples. To verify the effectiveness of our method, we applied the proposed method to the signals measured by FCS. The experimental results showed that the proposed method could automatically factorize the signals and could succeed to obtain the similar results with the manual investigations.

2. Method

In FCS, autocorrelation function (ACF) was extracted from the given time series signals measured from a living cell. They are represented as a feature vector. ACF may include several components related with different origins. Usually a set of feature vectors is obtained by measuring ACF from different cells in the same situation. The set of feature vectors is represented as a matrix. To analyze the protein dynamics of such cells, we have to decompose the matrix into the components (the basis vectors). The basis vectors can be modeled by the probability density function of the Boltzmann distribution. Usually they are modeled by fitting a model function using the nonlinear least squares method. Since both the ACFs and the basis vectors are non-negative, we have to decompose the matrix with the non-negative coefficients.

Non-negative matrix factorization (NMF) [11], [12] was proposed to decompose a given non-negative matrix into a non-negative basis matrix and a coefficient matrix. In this paper, we combine the idea of this non-negative decomposition with the nonlinear least squares fitting of a model function to the basis vectors. Once the basis vectors are modeled by the model function, we can estimate the diffusion time of each component and the component ratios from the estimated probability densities. For example, the diffusion time corresponding to a component can be calculated from its Boltzmann distribution.

2.1 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 the 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 the 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.

Figure 1 shows the schematic graph of the energy distribution of several particles. Here $P(I)$ is the total probability of the energy distribution and $P(I_i)$ is the probability of energy distribution on the i -th component (namely i -th system). We can assume that each particle in the system independently exists and the energy distribution of the particles can be represented by the Boltzmann distribution law when the dynamics of particles are measured in thermal equilibrium and its density in this system is sufficiently low. This situation is shown in Figure 1 (a). Here the i -th system represents the dynamics of the particle i .

From the Boltzmann distribution law, the number of particles N_j^i at the energy level ε_j in the i -th system is defined as follows:

$$N_j^i = A_i \, exp\left(-\frac{\varDelta\varepsilon_j}{k_B T_i}\right) \tag{1}$$

where A_i denotes a constant of i -th system that is determined from the number of particles in the lowest energy level and the statistical weights at this energy level. The quantity $\varDelta\varepsilon_j$ represents the difference of the energy level ε_j and the lowest energy level. Two parameters k_B and T_i are the Boltzmann constant and the absolute temperature of this system respectively.

The denominator $k_B T_i$ means the quantity of heat in the i -th system, therefore $k_B T_i$ must be proportional to the energy E_i in this system. Then the energy E_i of the particle i can be represented as follows:

$$E_i = \frac{1}{2} m_i {v_i}^2 \tag{2}$$

where m_i is the mass of the particle i and v_i is the velocity of the particle i in the measurement volume. When the measurement volume is sufficiently small, the effect of v_i becomes almost negligible and the energy E_i is proportional to the mass m_i . The diffusion time τ_i of the particle i is also proportional to the mass m_i . So $k_B T_i \propto \tau_i$ and $\varDelta\varepsilon_j \propto \left(\tau_j^i - \tau_0^i\right)$. When we define the time interval as $\Delta t_j = \tau_j^i - \tau_0^i$, the number of the particles N_j^i can be expressed as follows:

$$N_j^i \approx A_i \, exp\left(-\frac{\Delta t_j}{\tau_i}\right) \tag{3}$$

where the time interval Δt_j is the same in all systems because the diffusion time of the lowest energy level in the i -th system τ_0^i is independent in the particle dynamics.

The total probability of energy distribution $P(\Delta t_j \mid I)$ is defined as follows:

$$P(\Delta t_j \mid I) = \sum_i P(\Delta t_j \mid I_i) \tag{4}$$

where $P(\Delta t_j \mid I_i)$ is the probability of energy distribution in the i -th system and can be derived from the number of particles N_j^i . From these reasons, the total probability $P(\Delta t_j \mid I)$ can be defined as follows:

$$P(\Delta t_j \mid I) = \sum_i P(\Delta t_j \mid I_i) = \sum_i a_i \, exp\left(-\frac{\Delta t_j}{\tau_i}\right) \tag{5}$$

where a_i is the amplitude of the i -th system. This means that the total probability can be expressed as a linear combination of non-negative components.

Normalized ACF of the time series signal is equivalent to the total probability $P(\Delta t_j \mid I)$ because ACF is calculated from the measurement intensities with the time difference Δt_j .

2.2 Fluorescence Correlation Spectroscopy

FCS is one of the techniques to measure the fluorescence intensity fluctuations caused by fluorescent probe movement of free diffusion and to estimate diffusion times and existence ratios of fluorescent probes from autocorrelation function (ACF) calculated from the fluorescence intensity fluctuations. ACF is defined as follows:

$$G(\tau)=\frac{\langle I_t I_{t+\tau}\rangle}{\langle I\rangle^2} \tag{6}$$

where I_t is the signal intensity at time t . The diffusion time τ is defined as $\tau = \Delta t$. The quantity $\langle \mathrm{I}\rangle^2$ is the square of the averaged signal intensity.

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

$$G(\tau)=1+\frac{1}{N}\sum_i F_i\left(1+\frac{\tau}{\tau_i}\right)^{-1}\left(1+\frac{\tau}{s^2\tau_i}\right)^{-1/2} \tag{7}$$

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 $G(0)$) is determined by the reciprocal of the number of fluorescence molecules in detection volume. To calibrate the measurement device of FCS, ACFs of rhodamine 6G (Rh6G) water solution were measured for 30s five times at 10s interval, then the diffusion time (τ_{Rh6G}) and the structure parameter s were obtained by one-component fitting of the measured ACF in each sample.

Usually ACFs are obtained from different cells in the same situation and the statistical properties are investigated.

2.3 Factorization method

To analyze the protein dynamics of many cells, we have to decompose the matrix of ACFs into the components (the basis vectors). Since both the ACFs and the basis vectors are non-negative, we have to decompose the matrix with the non-negative coefficients.

Non-negative matrix factorization (NMF) [11], [12] was proposed to decompose a given non-negative matrix into a non-negative basis matrix and a coefficient matrix. We combine this non-negative decomposition with the nonlinear least squares fitting of a model function.

NMF decomposes the given $n\times m$ input matrix $\boldsymbol{V}$ into a $n\times r$ basis matrix $\boldsymbol{W}$ and an $r\times m$ coefficient matrix $\boldsymbol{H}$ as follows:

$$\boldsymbol{V} \approx \boldsymbol{WH} \tag{8}$$

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

NMF uses the divergence of $\boldsymbol{V}$ from $\boldsymbol{WH}$ as the measure of the cost for factorization. The objective function in NMF is given as follows:

$$D(\boldsymbol{V}\,||\,\boldsymbol{WH})=\sum_{ij}\left(\boldsymbol{V}_{ij} log\frac{\boldsymbol{V}_{ij}}{(\boldsymbol{WH})_{ij}}-\boldsymbol{V}_{ij}+(\boldsymbol{WH})_{ij}\right) \tag{9}$$

From this objective function (9), the multiplicative update rules of the basis and coefficient matrices in NMF were derived as follows:

$$\boldsymbol{W}_{ia} \leftarrow \boldsymbol{W}_{ia}\sum_{\mu}\frac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}}\boldsymbol{H}_{a\mu} \tag{10}$$

$$\boldsymbol{W}_{ia} \leftarrow \frac{\boldsymbol{W}_{ia}}{\sum_j \boldsymbol{W}_{ja}} \tag{11}$$

$$\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \sum_i \boldsymbol{W}_{ia} \frac{\boldsymbol{V}_{i\mu}}{(\boldsymbol{WH})_{i\mu}} \tag{12}$$

The proof of these multiplicative update rules are shown in [12]. Initial values of $\boldsymbol{W}$ are usually randomly assigned. In the following experiments, all random values were generated using Mersenne Twister algorithm (mt19937ar.c).

Since it is not guaranteed to reflect a physical phenomenon in the basis matrix obtained using NMF, we have to modify the basis vectors by fitting a model function.

In the current FCS, usually the model shown in equation (7) is fitted to the measured ACFs. But it is known that molecular dynamics in thermal equilibrium follows the Boltzmann distribution. An exponential function such as the Boltzmann distribution is often used in spectroscopy but it is not commonly used for the analysis of FCS [8]. To modify the original NMF, the probability density functions of the Boltzmann distribution are fitted to the basis vectors of r-th rank $\boldsymbol{w}_r$ by using the nonlinear least square method. From equation (5), the probability density function of the Boltzmann distribution is given as follows:

$$\boldsymbol{W}_{ir} = \boldsymbol{A}_r \, exp\left(-\frac{t_i}{\boldsymbol{\tau}_r}\right) \tag{13}$$

This fitting process is repeated at each step in the iterations of the NMF update.

3. Results

We applied the proposed method to two kinds of FCS data that were measured from the fluorescent molecule in water solution and the functional protein in living cell. In water solution data, the fluorescent fluctuations of Rh6G were used as a standard sample. In living cell data, we used Signal transducers and activators of transcription 3 (STAT3). The fluorescent fluctuations of functional protein were fused to the enhanced green fluorescence protein (EGFP). STAT3 has been shown to play pivotal roles in the cytokine signaling pathway, and also in regulating cell growth and differentiation. STAT3 is activated by stimulation with interleukin-6 (IL-6) which is a multifunctional cytokine. Molecular weight of STAT3 changes from monomer to dimer after IL-6 stimulation. In this paper, we used STAT3 measurement data in the nucleus before and after IL-6 stimulation because its diffusion time is expected to change into slow diffuse.

3.1 Analysis of energy distribution

We applied the proposed automatic factorization method to the 54 samples of Rh6G data that were measured on a 10^{-7} M concentrated solution. The 142×54 input matrix $\boldsymbol{V}$ was obtained by using these 54 samples. The number of basis vector must be one because Rh6G has only one component. The proposed method was applied to this data. Figure 2 shows the approximation of $\boldsymbol{v}$ by $\boldsymbol{wh}$, the products of the basis vector $\boldsymbol{w}$ and the coefficients of each sample $\boldsymbol{h}$. Here the basis vector $\boldsymbol{w}$ was approximated by fitting the model function shown in equation (13). This suggests that our proposed method gives a good fitting except in slow diffusion times.

Table 1 summarizes the diffusion times of Rh6G that were estimated by the manual analysis and by the proposed method. The manually estimated diffusion time was 24.9μs when it was calculated as the average of the 54 samples. The standard deviation of this diffusion time was 11.5μs. The diffusion time estimated by fitting the model function to the basis vector $\boldsymbol{w}$ was 39.0μs. We can say that the estimated diffusion time seems biologically valid.

Table 1. Estimated Diffusion Time of Rh6G

Using method	Diffusion time / μs (ratio / %)
FCS manual analysis method	24.9 ±11.5 (100)
The proposed method	39.0 (100)

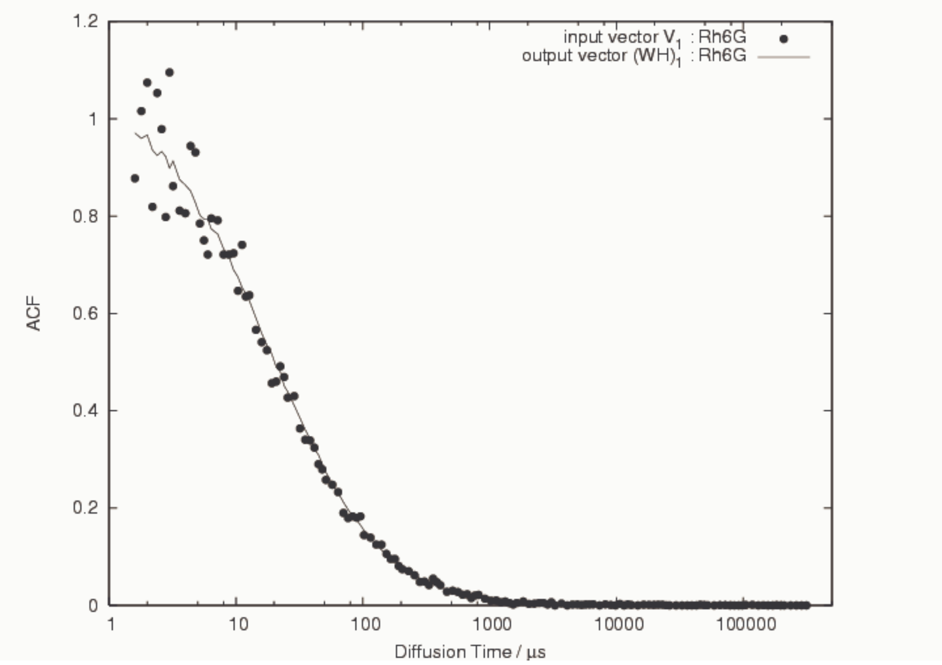

Figure 2. Automatic factorization of Rh6G data measured by FCS. FCS measurements were carried out in water solution. The closed circles show the samples measured by FCS and the line is the result of approximation of the model function by using the proposed method.

3.2 Results for STAT3 data

STAT3 was fused to EGFP (STAT3-GFP) and the 47 samples and the 43 samples before and after IL-6 stimulation were measured by using FCS [2]. We had the 124×47 input matrix $\boldsymbol{V}$ for before IL-6 stimulation and the 127×43 input matrix $\boldsymbol{V}$ for after IL-6 stimulation. For each input matrix the proposed factorization method was applied. For this data, we assumed the number of basis vectors, namely rank of the NMF, was at most three because STAT3 in the nucleus of living cell is inhibited free diffusion and exists as the monomeric form or the dimeric form before and after IL-6 stimulation, respectively.

The results of automatic factorization for STAT3-GFP measured by FCS before and after IL-6 stimulation were shown in Figure 3. The results for before IL-6 stimulation and after IL-6 stimulation are shown in Figure 3 A and B, respectively. The closed circles show the samples measured by FCS. Line is the result of the approximation by NMF-based automatic factorization. The open circles, squares and triangles are the estimated basis of each diffusion component 1, 2 and 3, respectively. These results are reasonable because the number of samples with faster diffusion times increase after the stimulation.

The distribution of the diffusion times of STAT3-GFP measured by FCS in the nucleus of living cell before and after IL-6 stimulation is shown in A and B of Figure 4, respectively. The manually estimated diffusion times of each measurement are shown in the scatter plots of open diamonds. Bars shows the diffusion times calculated from the estimated basis vectors by the proposed method. The distribution of the diffusion times and the existence ratios are shown in Figure 4. The diffusion time of the main component obtained by the automatic factorization is 702.1μs (48.7%) and the other components are 3830.5μs (26.3%) and 2385.8μs (25.1%) as shown in Figure 4 A. On the other hand, the diffusion time of the main component for after stimulation is 831.4μs (94.7%) and the other components are 4876.4μs (2.70%) and 2994.4μs (2.63%) as shown in Figure 4 B. The diffusion time of the main component increased after IL-6 stimulation. This reflects the physical phenomenon that changes from the monomeric form to the dimeric form. These results show the validity of the proposed method.

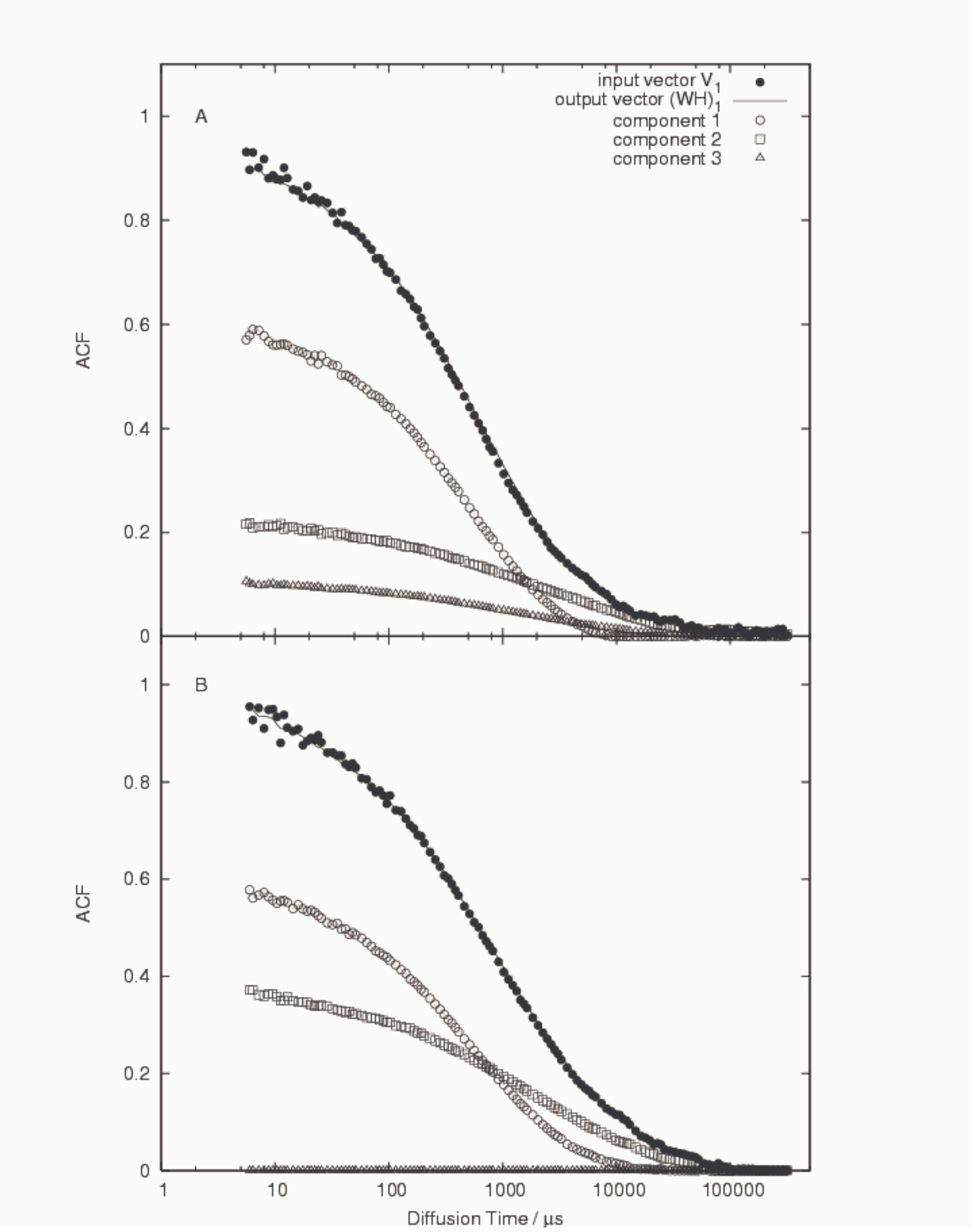

Figure 3. Automatic factorization of STAT3-GFP measured by FCS before and after IL-6 stimulation. FCS measurements were carried out for STAT3-GFP in the nucleus of living cell. Normalized ACF before and after IL-6 stimulation is shown in A and B, respectively. The closed circles show the samples measured by FCS (A, B). Line is the result of the approximation by NMF-based automatic factorization (A, B). The open circles, squares and triangles are the estimated basis vectors of each diffusion component 1, 2 and 3, respectively (A, B).

4. Discussion

The proposed method gave the similar tendency with the biological knowledge. In General, the current biological knowledge about the state of STAT3 in the nucleus is as follows. Before IL-6 stimulation, the main component of STAT3 exists as monomer and the sub components exist as slower movements. However, after IL-6 stimulation, a main component of STAT3 exists as dimer. Such a biological knowledge was confirmed by using classical biological experimental methods in dead cell and was also verified by using FCS in living cell [2].

In our experimental results, the different diffusion time of the main components were estimated by using our proposed factorization method before and after IL-6 stimulation. The results of the main components are probably STAT3 monomer and dimer. The other diffusion times of the sub components were over 2000µs before

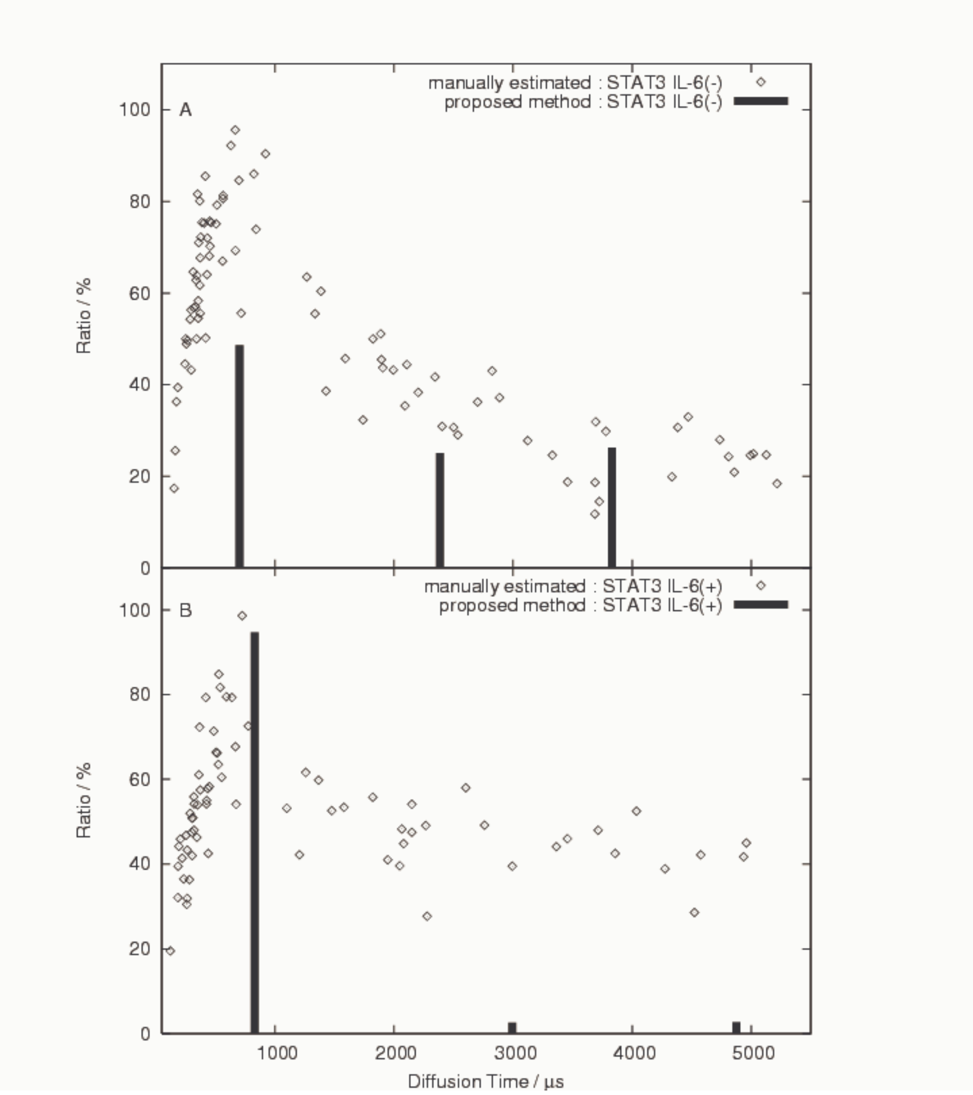

Figure 4. The distribution of the diffusion times of STAT3-GFP measured by FCS in the nucleus of living cell before and after IL-6 stimulation is shown A and B, respectively. The manually estimated diffusion times of each measurement are shown in the scatter plots of open diamonds. Bars shows the diffusion times calculated from the estimated basis vectors by the proposed method.

IL-6 stimulation. These results of sub components may be inhibited by the free diffuse of STAT3. These results have the similar tendency of the biological knowledge. The proposed method can also give the same results with the ordinal method in FCS data analysis (Figure 4). Even if our proposed method could not obtain the results of completely same tendency, it may be caused for a spectroscopy problem such as the effect of triplet state. This problem can be solved by changing the current model function to another one.

In ordinal FCS data analysis, the diffusion times and the existence ratios are estimated by fitting the equation (7) to each measurement sample. When we need a statistics that reflects the physical phenomena measured by FCS, we have to manually analyze the diffusion times. In this manual treatment of the data, there is a possibility to have danger that the subjectivity of researchers is included. The manual analysis requires a great labor because the analysis has to perform for each sample. However, the proposed method makes automatic statistical analysis of all samples possible. From these reasons, the proposed method is useful.

For future works, we have to modify NMF to introduce the probability density function of the Boltzmann distribution into the multiplicative update rules. This modified NMF will be verified by using the simple simulation data that is generated by the model function. Also we have to select the number of basis vectors automatically. We will try to use model selection techniques such as cross-validation. Thereafter we have to confirm the effectiveness of the proposed method by applying to other biological data sets.

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 provision of biological materials and the measurement of FCS.

References

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

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

[3] Kitamura, A., Kubota, H., Pack, C.-G., Matsumoto, G., Hirayama, S., Takahashi, Y., Kimura, H., Kinjo, M., Morimoto, R.I., Nagata K.: Cytosolic chaperonin prevents polyglutamine toxicity with altering the aggregation state. Nature Cell. Biol. 8 (2006) 1163-1170

[4] Ehrenberg, M., Rigler, R.: Rotational brownian motion and fluorescence intensify fluctuations. Chem. Phys. 4 (1974) 390-401

[5] Elson, E.L., Magde, D.: Fluorescence correlation spectroscopy. I. Conceptual basis and theory. Biopolymers 13 (1974) 1-27

[6] Koppel, D.E.: Statistical accuracy in fluorescence correlation spectroscopy. Phys. Rev. A 10 (1974) 1938-1945

[7] Rao, R., Langoju, R., Go1sch, M., Rigler, P., Serov, A., Lasser, T.: Stochastic Approach to Data Analysis in Fluorescence Correlation Spectroscopy. J. Phys. Chem. A 110 (2006) 10674-10682

[8] Kim, H.D., Nienhaus, G.U., Ha, T., Orr, J.W., Williamson, J.R., Chu, S.: Mg2+-dependent conformational change of RNA studied by fluorescence correlation and FRET on immobilized single molecules. Proc. Natl. Acad. Sci. USA 99 (2002) 4284-4289

[9] Comon, P.: Independent component analysis, A new concept? Signal Processing 36 (1994) 287-314

[10] Delorme, A., Sejnowski, T., Makeig, S.: Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage 34 (2007) 1443–1449

[11] Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401 (1999) 788-791

[12] Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. Adv. Neural Info. Proc. Syst. 13 (2001) 556-562

Kenji Watanabe graduated from the Hokkaido University in 2002 and received the M.MedSc. degree from Hokkaido University in 2004. He is currently a student of Dr. Kurita in University of Tsukuba.

Takio Kurita graduated from Nagoya Institute of Technology in 1981 and received the Dr.Eng. degree from University of Tsukuba in 1993. He joined the Electrotechnical Laboratory, AIST, MITI, Japan in 1981. From 1990 to 1991 he was a visiting research scientist at Institute for Information Technology, NRC, Ottawa, Canada. He is currently Deputy Director of Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology (AIST). His current research interests include statistical pattern recognition and neural networks.
(Home page: http://staff.aist.go.jp/takio-kurita/)