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

Kenji Watanabe[1] and Takio Kurita[2]

[1] 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
[2] National Institute of Advanced Industrial Science and Technology (AIST),
AIST Central 2, 1-1-1 Umezono, Tsukuba-shi, Ibaraki-ken, 305-8568 Japan
{kenji-watanabe, takio-kurita}@aist.go.jp

Abstract. We proposed automatic factorization method of 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 represented by model functions and the signals are factorized as the non-negative sum of the model functions. Analytical accuracy of our proposed method was verified by using biological data that were measured by FCS. The experimental results showed that our method could automatically factorize the signals and the obtained components were similar with the ones obtained manually.

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 be more important in the future. However, the current methods of time series analysis are 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, there is a possibility to have danger that the subjectivity of researchers is included in the results obtained by the current methods because the examination of analytical results and judgments of re-analysis are decided manually. To improve the current methods, 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 the statistical analysis of the large amount of samples is required to estimate the components. In molecular biology, the components are manually found through statistical investigation.

M. Ishikawa et al. (Eds.): ICONIP 2007, Part II, LNCS 4985, pp. 798 – 806, 2008.

Automatic signal factorization has been examined in a lot of fields, for example, factor analysis, independent component analysis (ICA) [9], [10], non-negative matrix factorization (NMF) [11], [12]. Especially, NMF is probably effective for the factorization of non-negative energy distribution such as a molecular dynamics in thermal equilibrium. On the other hand, ICA is not suitable for this application because the independency is not guaranteed.

In this paper, we proposed 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 model functions derived by considering its physical phenomena and is fitted by the nonlinear least squares method. By using NMF approach, we can directly find the components included in the autocorrelation functions from the all samples.

To verify the effectiveness of our method, we applied the proposed method to the signals obtained by FCS.

2 Method

In FCS, autocorrelation function (ACF) was extracted from time series signals measured from a living cell and 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 Boltzmann distribution law. 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. We combine this non-negative decomposition with the nonlinear least squares fitting of model function. 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 basis vector can be calculated from the probability density function estimated for the basis vector considering its Boltzmann distribution.

2.1 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 deduce diffusion times and existence ratios of fluorescent probes from autocorrelation function (ACF) calculated from the fluorescence intensity fluctuations. ACF is defined as follow:

$$G(\tau)=\frac{\langle \mathbf{I}_t \mathbf{I}_{t+\tau} \rangle}{\langle I \rangle^2} \tag{1}$$

where $\mathbf{I}_t$ is the signal intensity in time t. Diffusion time τ is defined as $\tau = \Delta t$. $\langle I \rangle^2$ is 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:

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

where F_i and τ_i are the fraction and diffusion time of component i, respectively, $\mathbf{N}$ 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. ACF of rhodamine 6G (Rh6G) 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.

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

2.2 Signal Factorization

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 model function.

NMF decomposes the given $n \times m$ input matrix V into a $n \times r$ basis matrix W and an $r \times m$ coefficient matrix as follow:

$$V \approx WH \tag{3}$$

This means that WH is an approximation of the matrix V.

NMF uses the objective function that is the divergence of V from WH as the measure of the cost for factorization. The objective function in NMF is given as follow:

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

From this objective function (4), we can derive multiplicative update rules in NMF as follow:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \tag{5}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}} \tag{6}$$

$$H_{a\mu} \leftarrow H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}} \tag{7}$$

The proof of these objective function and 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).

There is no guarantee to reflect a physical phenomenon in the basis matrix computed using NMF.

In FCS, generally ACF are fitted by using equation (2). But molecular dynamics in thermal equilibrium follows Boltzmann distribution law. An exponential function that is represented as like Boltzmann distribution is often used in spectroscopy but the exponential function is uncommonly used to the analysis of FCS [8]. To modify the original NMF, the probability density functions of Boltzmann distribution law are fitted to the basis vectors $\boldsymbol{w}_r$ by using the nonlinear least square method. The probability density function of Boltzmann distribution law is given as follow:

$$w_r = \mathbf{A}\, exp\left(-\frac{\tau}{\tau_r}\right) \tag{8}$$

where $\mathbf{A}$ is amplitude and τ_r is the diffusion time of $\boldsymbol{w}_r$. This fitting process is repeated at each iteration of the NMF update.

3 Experiments and 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 Results for Rh6G Data

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 $\mathbf{142 \times 54}$ 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. The approximation of $\boldsymbol{V}$ by $\mathbf{w}\mathbf{h}^T$, the products of the basis vector $\mathbf{w}$ and the coefficients of each sample $\mathbf{h}$, is shown in Fig. 1. Here the basis vector $\mathbf{w}$ was approximated by fitting the model function shown in equation (8). This suggests that our proposed method gives a good fitting except in slow diffusion times.

Table 1 shows that the diffusion times of Rh6G that were estimated manually and by our proposed method. The manually estimated diffusion time was 24.9 μ s when it was calculated as the average of the 54 samples. The standard deviation of this diffusion

Table 1. Estimated Diffusion time of Rh6G

Using method	Diffusion time / μ s (ratio / %)
Manually estimated	24.9 (100)
Proposed method	39.0 (100)

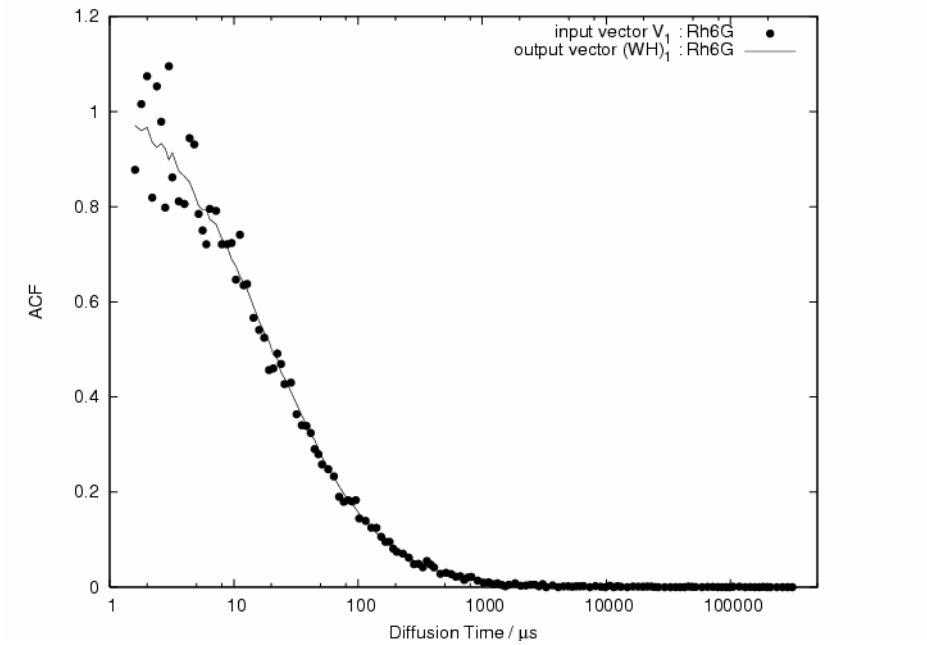

Fig. 1. Automatic factorization of Rh6G data measured by FCS. FCS measurements were carried out in water solution. The closed circles show the samples measured by FCS and the line is the result of approximation.

time was 11.5. The diffusion time estimated by fitting the model function to the basis vector $\mathbf{w}$ was 39.0 μ s. We can say that the estimated diffusion time seems biologically valid.

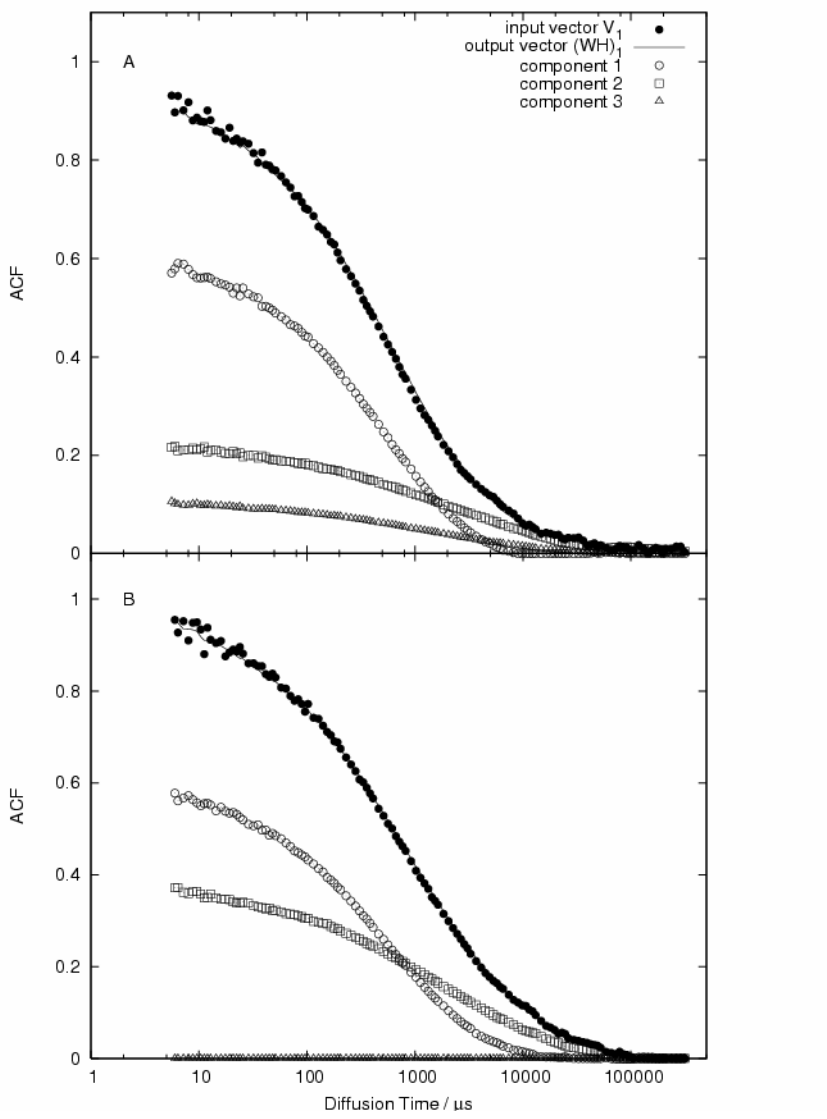

Fig. 2. Automatic factorization of STAT3-GFP measured by FCS before and after IL-6 stimulation. FCS measurements were carried out for STAT3-GFP in the nucleus of living cell. Normalized ACF before and after IL-6 stimulation is shown A and B, respectively. The closed circles show the samples measured 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 of each diffusion component 1, 2 and 3, respectively (A, B).

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]. Thus, we can obtain the $\mathbf{124}\times\mathbf{47}$ input matrix $\boldsymbol{V}$ for before IL-6 stimulation and the $\mathbf{127}\times\mathbf{43}$ 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,

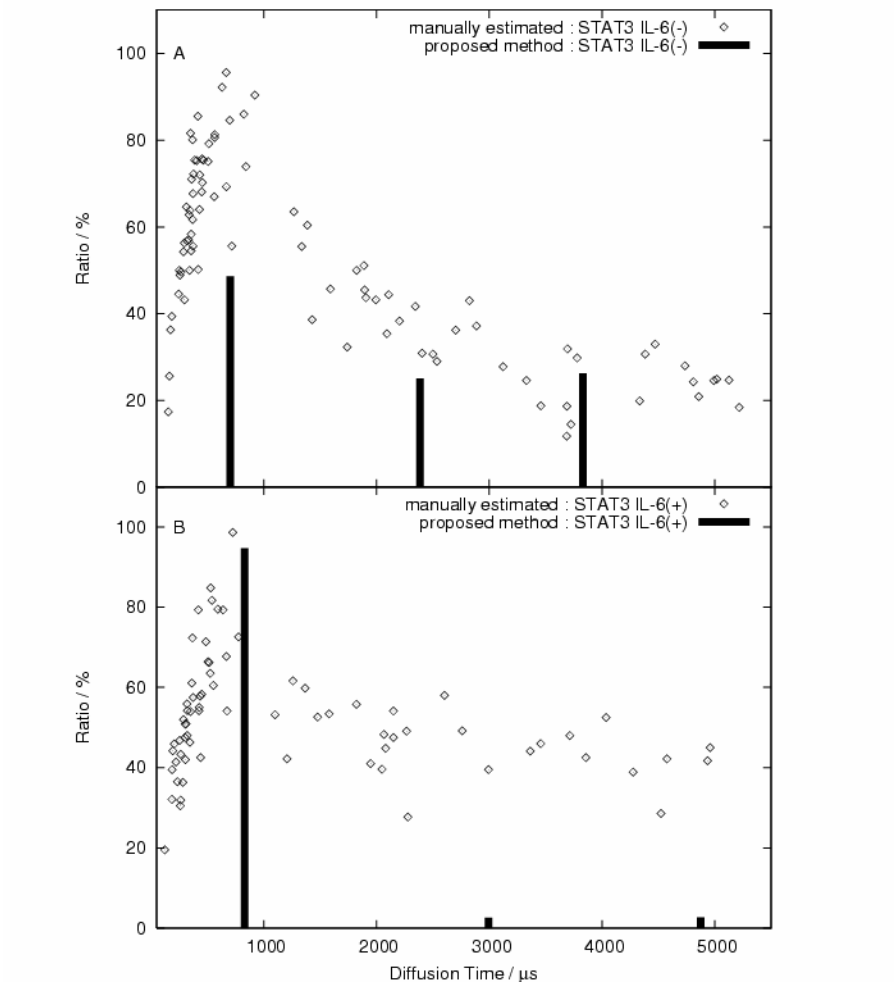

Fig. 3. The distribution of the diffusion times of STAT3-GFP measured by FCS in the nucleus of living cell before and after IL-6 stimulation is shown A and B, respectively. The manually estimated diffusion times of each measurement are shown in the scatter plots of open diamonds. Bars shows the diffusion times calculated from the estimated basis vectors by the proposed method.

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 Fig. 2. Fig.2 A and B show the results for before IL-6 stimulation and after IL-6 stimulation, 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 Fig. 3, 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 Fig. 3. The diffusion time of the main component obtained by the automatic factorization is 702.1 μ s (48.7%) and the other components are 3830.5 μ s (26.3%) and 2385.8 μ s (25.1%) as shown in Fig. 3 A. On the other hand, the diffusion time of the main component for after stimulation is 831.4 μ s (94.7%) and the other components are 4876.4 μ s (2.70%) and 2994.4 μ s (2.63%) as shown in Fig. 3 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.

4 Discussion

The proposed method gave the similar tendency with the previous biological theory. In General, the current biological theory about the state of STAT3 in the nucleus is as follow. Before IL-6 stimulation, the main component of STAT3 exists as monomer and the sub components exist as lower movements. However, after IL-6 stimulation, a main component of STAT3 exists as dimer. Such a biological theory 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 2000 μ s before 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 theory. The proposed method can also give the same results with the ordinal method in FCS data analysis (Fig 3). 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 model function to another.

In ordinal FCS data analysis, the diffusion times and the existence ratios are estimated by fitting the equation (2) 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 Boltzmann distribution law to 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. Thereafter we have to confirm the effectiveness of the proposed method by applying to other biological data sets.

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, 164–171 (2006)
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, 1264–1273 (2004)
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, 1163–1170 (2006)
4. Ehrenberg, M., Rigler, R.: Rotational brownian motion and fluorescence intensify fluctuations. Chem. Phys. 4, 390–401 (1974)
5. Elson, E.L., Magde, D.: Fluorescence correlation spectroscopy. I. Conceptual basis and theory. Biopolymers 13, 1–27 (1974)
6. Koppel, D.E.: Statistical accuracy in fluorescence correlation spectroscopy. Phys. Rev. A 10, 1938–1945 (1974)
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, 10674–10682 (2006)
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, 4284–4289 (2002)
9. Comon, P.: Independent component analysis, A new concept? Signal Processing 36, 287–314 (1994)
10. Delorme, A., Sejnowski, T., Makeig, S.: Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage 34, 1443–1449 (2007)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
12. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. Adv. Neural Info. Proc. Syst. 13, 556–562 (2001)