Research article

Evaluation of protein crystallization state by sequential image classification

Kuniaki Kawabata
Kawabata Intelligent System Research Unit, RIKEN (The Institute of Physical and Chemical Research), Wako, Japan

Kanako Saitoh
Kawabata Intelligent System Research Unit, RIKEN (The Institute of Physical and Chemical Research),
Wako, Japan and
Saitama University, Saitama, Japan

Mutsunori Takahashi
Saitama University, Saitama, Japan

Hajime Asama
The University of Tokyo, Kashiwanoha, Japan

Taketoshi Mishima
Saitama University, Saitama, Japan, and

Mitsuaki Sugahara and Masashi Miyano
SPring-8 Center, Harima Institute, RIKEN (The Institute of Physical and Chemical Research), Sayo, Japan

Abstract

Purpose – The purpose of this paper is to present classification schemes for the crystallization state of proteins utilizing image processing.
Design/methodology/approach – Two classification schemes shown here are combined sequentially.
Findings – The correct ratio of experimental result using the method presented here is approximately 70 per cent.
Originality/value – The paper is a contribution to automated evaluation crystal growth, combining two classifiers based on specific visual feature, sequentially.

Keywords Crystallization, States of matter, Image processing, Multivariate analysis

Paper type Research paper

1. Introduction

Protein crystallization takes a great deal of time, and to achieve more rapid progress in structural genomics research it is necessary to develop an efficient structural analysis process. Automated systems have been developed (Sugahara *et al.*, 2002) for certain processes (dispensation of precipitant solutions, photography of protein solution and supervising a large number of protein solution samples), yet there remains a great need for an automated classification procedure for protein crystallization states. To this end, our group began investigating a method for classifying protein crystallization states by image processing. Two discrimination methods (Saitoh *et al.*, 2005; Kawabata *et al.*, 2006) have been presented to date, by which the crystallization droplets are categorized into five groups, and then droplets containing amorphase or crystalline objects into two groups. It is important to evaluate its performance by combining them because there is the deference in the visual features between early stage and latter stage of crystallization growth state. For example, the image of categories 0-3 does not include shape features but the one of categories 4-9 contains shape features as visual impression.

In the present study, these two categorization methods are applied sequentially, providing classification into six protein crystallization states.

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0260-2288.htm

Sensor Review
28/3 (2008) 242–247
© Emerald Group Publishing Limited [ISSN 0260-2288]
[DOI 10.1108/02602280810882607]

This work has been partially supported by a research grant of the Okawa Foundation for Information and Telecommunication (05-23) and Takeda-rika Co.

2. Crystallization state classification

2.1 Target image, categorization and pre-processing

Figure 1 shows photographs of typical protein crystallization, obtained using the TERA system developed by RIKEN (Sugahara *et al.*, 2002). The original image was captured at 40 × magnification and shows the entire protein solution sample from above.

Although our ultimate goal is to realise a totally automated evaluation system for the ten categories, we attempt to evaluate categories 0-3 as the first step. The states from classes 0 to 3 do not have shape, while those from 4 to 9 have some shape, for example; circular form, needle and plate, etc. Therefore, to evaluate samples in categories 0-3 and 4-9, we think that features based on image pattern and based on shape should be effective, respectively. We attempt to evaluate these two sets of categories sequentially. Our target classification, which is a reclassification of the RIKEN system, is as shown in Figure 1. In the RIKEN, the samples in clear and precipitate stages (0-3) occupy approximately 70 per cent of all supervised samples, so we consider that our target categorization of Step 1 is very effective for categorization work. Group E includes non-crystalline objects (Group 4 of the RIKEN categorization) with crystalline objects. For example, the objects in Groups 4-9 have circular shape, so it may be possible to classify them by measuring the basic shape feature. This subject is handled in second step categorization.

As local features may not be resolved using the original images (1,392 × 1,040 pixels), the original images were divided into small areas (150 × 150 pixels), and each sub-image was processed separately (Figure 2). The images are first converted to gray-scale images and enhanced by differential processing. Binarized processing is also performed for edge detection.

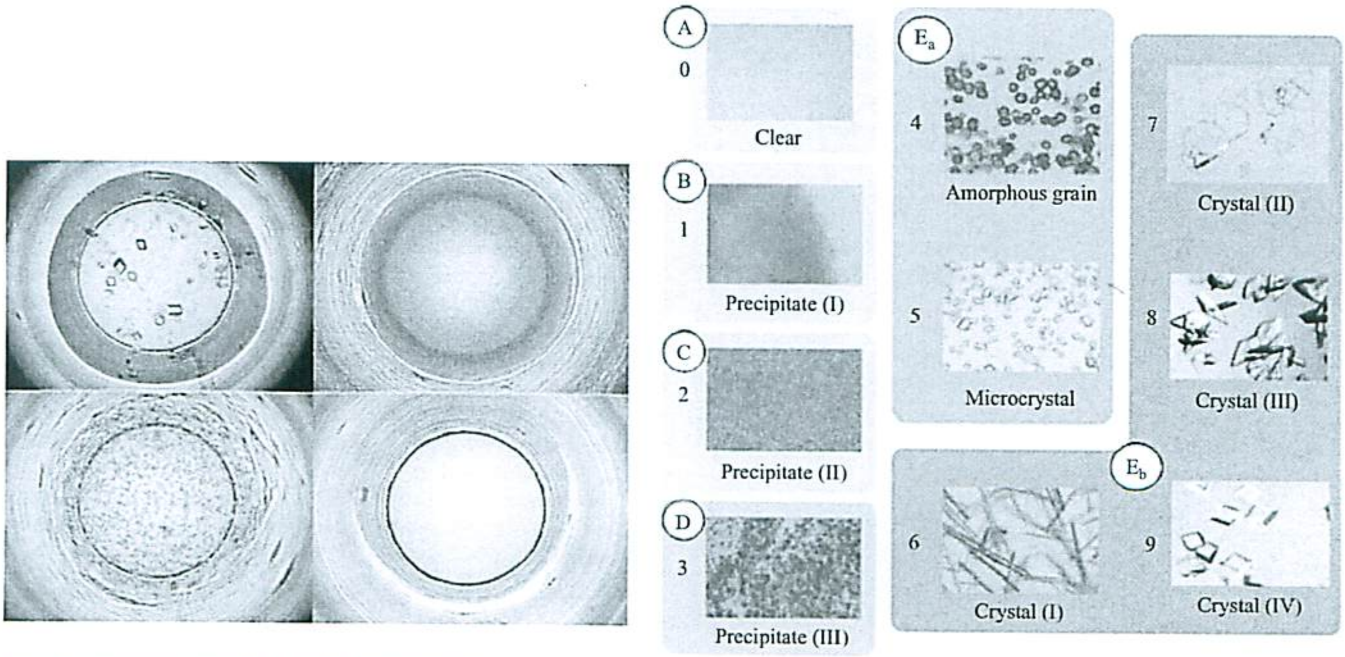
Initially, the original color images are transformed into 256 level gray-scale images, because color information of images is not utilized in this method. Next, a portion of the original image is manually extracted from inside the well. The processing object in this study is assumed to be only inside the well. The extraction size is 150 × 150 pixels, which is determined by considering approximate average size of microcrystals and crystals in the original images. Finally, the extracted image is differentiated with a sobel first-order differential filter. This process highlights the characteristic pattern of the image.

First step discrimination involves extracting features from both the unprocessed and differential enhanced images. It means that both of differentiated and non-differentiated images are utilized in first step (Figure 3). In second step discrimination, line segment features are extracted using the binarized images (Figure 4) based on the threshold value which is determined by statistical analysis (Kawabata *et al.*, 2006). The edge strength distribution of the images that do not include any crystalline object after Sobel filter operation is examined and utilized as an index for the threshold configuration. Specifically, the edge strength distribution of 100 images with a score of 0 (clear) (Figure 5) was investigated and it was found that a threshold value of 29 binarizes 99.9 per cent of the pixels of the whole distribution as background. This threshold is therefore employed to extract the contour lines.

2.2 Step 1: linear discrimination utilizing texture information

The images of the protein solution are first categorized by a textural analysis method. Textural feature values are calculated using a gray-level co-occurrence matrix (Haralick *et al.*, 1973), which is a popular statistical method in textural analysis. The co-occurrence matrix is defined in terms of the distances and angles among neighboring pixels, where each element of the matrix expresses the frequency that a pixel with a given gray level occurs in a certain spatial relationship with a pixel with another given gray level. The co-occurrence matrix for the present images consist of 14 feature values. Thus, each

Figure 1 Target images and ten categories (0-9) for protein crystal evaluation



Note: Target categories (A, B, C, D, E_a, E_b) are also shown

Figure 2 Pre-processing sequence for first step: (a) original image; the image size is 1,392 × 1,040 pixels; (b) 256-level greyscale image; (c) a portion of the original image; the image size is 150 × 150 pixels; (d) differentiated image

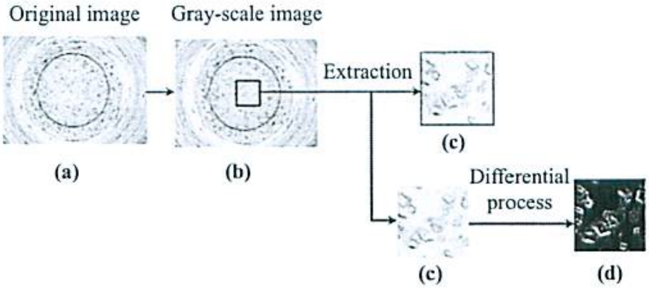
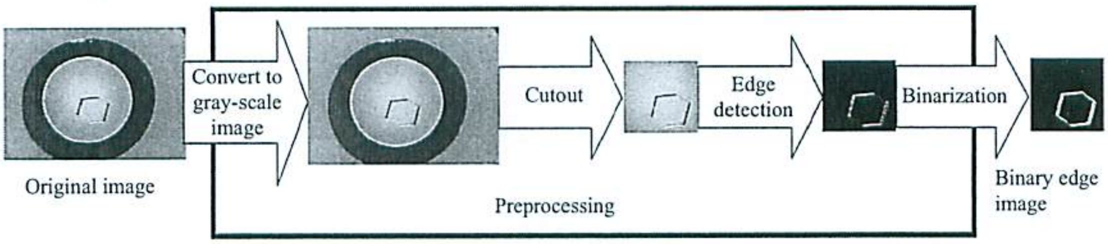


Figure 3 Pre-processing flow for second step



Notes: The flow consists of cutout, edge detection and binarization for binary edge image (size is 150 × 150 pixels)

Figure 4 Examples binarized images

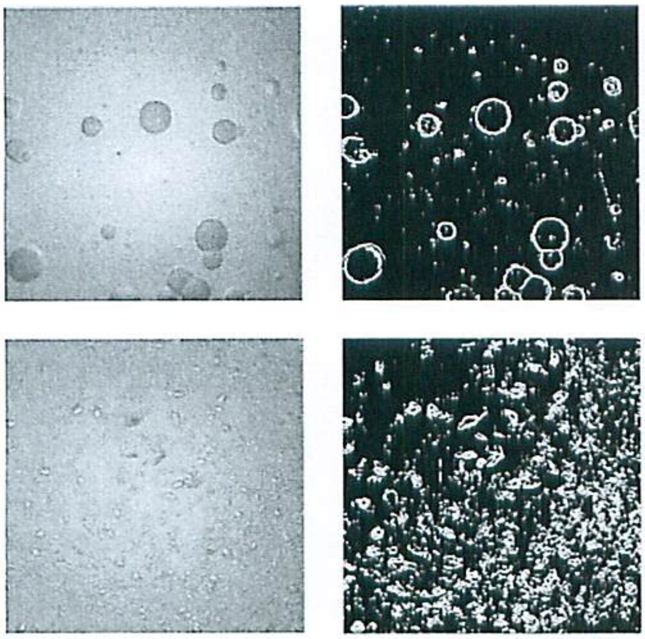
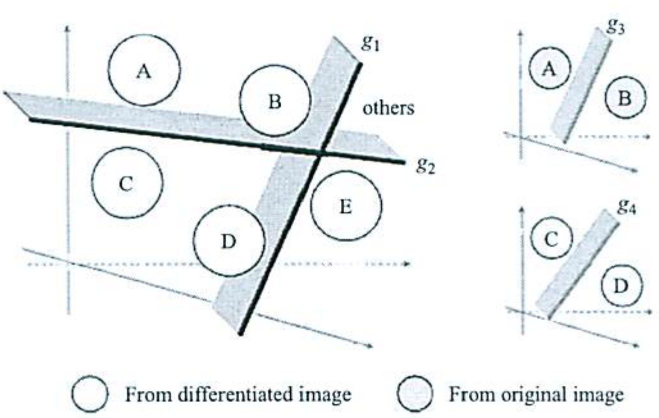


Figure 5 Discrimination procedure



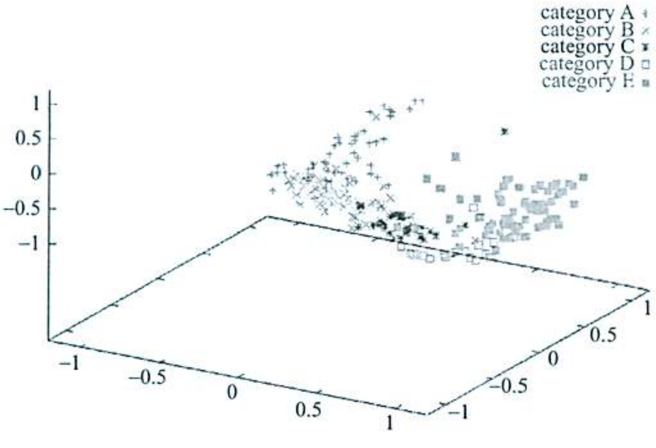
Notes: The procedure consists of two stages: (1) input data is classified into A, B or C, D or E,using the functions g_1 and g_2 and (2) the data are classified into A or B, or C or D using the functions g_3 and g_4)

image is expressed as a single point in 14 dimensional feature space. This step utilizes 150 × 150 pixel images, both unprocessed and after differential enhancement. Texture feature values (Haralick *et al.*, 1973) are calculated by using gray-level co-occurrence matrix. Each element of the matrix $\mathbf{P}_\delta$ (i, j) expresses the probability that is at once gray-scale value of one pixel is i and gray-scale value of another pixel located at φ direction and r pixels away from the former pixel is j. Where, the displacement between two pixels is denoted $\delta = (r, \varphi)$. The distance between pixels: r used to calculate the co-occurrence matrices is 1, and the direction: φ are 0°, 45°, 90°, 135°. The four directions correspond to horizontal, vertical and the two diagonals. By using the matrix $\mathbf{P}_\delta$ 14 texture feature are calculated (Haralick *et al.*, 1973).

Here, Figure 6 shows class mapping in 3D feature space by using principal component analysis (PCA). It can be confirmed that each class overlaps on the others. Especially, Group E formed a cluster in this feature space and it means that it is difficult to discriminate them into detail group (score 4-9) by using only the texture information. Therefore, as first step, five class categorization (Groups A-E) is experimented.

Using 14 kinds of extracted features, the crystallization state is classified into one of five categories in the first step (Saitoh *et al.*, 2005): clear (A), precipitate (i) (B), precipitate

Figure 6 Overview of class mapping in 3D feature space by using PCA



(ii) (C), precipitate (iii) (D), amorphous grains, microcrystals and crystals (E). Categorization is performed by linear discriminant analysis (Fisher, 1936), which is a standard technique of multivariate analysis (Fisher, 1936). Discriminant analysis divides the feature space into partial spaces.

To realize the five categories above, four linear discriminant functions (g_1, g_2, g_3, g_4) based on 874 sub-images are employed:

1 g_1 – classify the entire feature space into A, B, C, D or E;
2 g_2 – classify the entire feature space into (A, B) or (C, D, E);
3 g_3 – classify the partial space (A, B) into A or B; and
4 g_4 – classify the partial space (C, D) into C or D.

Functions g_1, g_2, and g_4 are performed using the feature values from differentiated images, and function g_3 is determined using the feature values from unprocessed images. Using functions g_1 and g_2, the entire group is classified into (A, B), (C, D) and E. Function g_3 is then applied to classify the space (A, B) into A and B, and g_4 classifies the space (C, D) into C and D.

2.3 Step 2: line feature extraction and discriminant analysis

In manual classification, samples in Category E (Figure 1) are separated into those suitable for X-ray diffraction analysis and those that are not on the basis of crystallinity. This is performed by observing the shapes of the objects in the crystallization solution, which can be regarded as characteristic contours for automated analysis. The line feature extraction and discriminant analysis method (Kawabata *et al.*, 2006) employed here uses the contours as feature values. The images in Group E are thus further categorized into the following three groups:

1 *Amorphous grains* (Type 4, group E_a): circular objects.
2 *Microcrystalline* (Type 5, group E_a): angular objects with length <0.05 mm, thickness < 0.01 mm, and width <0.01 mm.
3 *Crystalline* (Types 6–9, group E_b): angular objects with length ≥ 0.05 mm, thickness ≥ 0.01 mm, and width ≥ 0.01 mm.

Linear features are scanned within the binary edge images. The edge pixels that exist sequentially in-line are considered to be the line segments, and their lengths and numbers are determined. However, since it is impossible to predict the location where the crystalline objects will be formed in the crystallization solution and the location of the contour lines that will be obtained by pre-processing, the crystalline object images need to be scanned from every possible direction. Thus, raster scanning is conducted by altering θ, the scanning direction, (0–180°) as shown in Figure 7. An edge pixel that is detected on the scanning line is counted as the line segment, and the pixel is considered to be the origin of the n-th line segment and the scanning process is continued. The last edge pixel in a line-segment is considered to be the terminal and the length of the line segment is then determined.

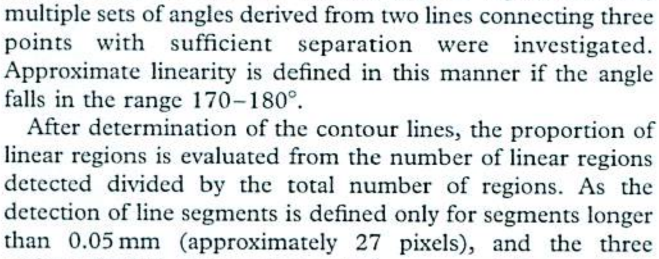
By this scanning process, the length of each line segment is calculated, and the total number of line segments is determined from the maximum length and the number of line segments. To extract the proportion of linear regions, the number of linear regions longer than 0.05 mm within the contour lines of the object is determined. It should be noted that the extracted contour lines may be somewhat distorted due to noise in pre-processing, although the lines may have originally been straight.

By observing the spatial relationships among pixels within a region, the linearity of the region can be determined. In the case that all pixels within a region are positioned in-line, if three arbitrary points within a region are selected, two linear curves connecting the center and two other points will essentially be oriented at 180°. If adjacent points are not selected, even slightly distorted linear lines will form a similar angle (180°). To evaluate whether a given region is linear, multiple sets of angles derived from two lines connecting three points with sufficient separation were investigated. Approximate linearity is defined in this manner if the angle falls in the range 170–180°.

After determination of the contour lines, the proportion of linear regions is evaluated from the number of linear regions detected divided by the total number of regions. As the detection of line segments is defined only for segments longer than 0.05 mm (approximately 27 pixels), and the three regions (width: three unit lengths) are regarded as a single combined region when evaluating the contour lines, the width of unit length was set at nine pixels. This processing step is performed using binarized 150×150 pixel sub-images, which differ from the image types used originally with these classification scheme (Kawabata *et al.*, 2006).

In the feature space consisting of maximum length, number of line segments, and proportion of linear regions, the variance between categories E_a (amorphous and microcrystalline) and E_b (crystalline) differs significantly (Table I, the data from 500 full-size images ($1,392 \times 1,040$ pixels) of each category).

Figure 7 Raster scanning the line segments on the cutout images in the direction θ



Note: The detected line segments information is utilized to evaluate the image

Table I Average and standard deviation of the group E_a and E_b

		E_a	E_b
Average	Maximum length	2,325,426	1,168,213
	Number of line segments	26.0	74.9
	Proportion of linear regions	0.05	0.20
Standard deviation	Maximum length	1,066,519	1,144,607
	Number of line segments	6.5	63.7
	Proportion of linear regions	0.08	0.14

The Mahalanobis (1936) generalized distance is thus employed as the criterion for discriminant analysis: g_5 – classify Group E into E_a or E_b.

Here, g_5 is detemined by using 200 images of Category A and 100 images of Category B (150 × 150 pixels).

The sequential classification scheme is shown in Figure 8. At first step, five class categorization (Groups A-E) from input data are done and at second step, two class categorization (group E_a and E_b) from Group E which is evaluated by first step. Finally, the input data are classified into six categories.

3. Classification results

The performance of sequential discrimination processing was evaluated by comparing the classification results with manual classification by a human expert. The accuracy is defined as the number of correct results divided by the total number of images categorized manually into a given group. A total of 874 TERA images annotated by a human expert at RIKEN were employed for evaluation, of which 435 were used as training images (A: 49, B: 45, C: 49, D: 49, E: 243). The remaining 439 images (A: 53, B: 71, C: 29, D: 41, E: 245) were used as the test set. The classification performance for Step 1 of the present method is shown in Table II the average of accuracy is 87.70 per cent.

The images classified into Group E were further evaluated by Step 2 processing, resulting in a total of six categories. Of the total 488 images classified as Group E, 243 were used as the training set (E_a: 95, E_b: 148), and the remaining 245 images (E_a: 125, E_b: 120) were used as the test set. Here, by our previous examinations utilizing full-size images (1,392 × 1,040 pixels) (Kawabata *et al.*, 2006), it is already confirmed that Step 2 method can classify them with 80.0 per cent accuracy.

The final classification performance after taking Step 2 of the sequential classification scheme is shown in Table III. The average of the accuracy is 65.38 per cent.

Figure 8 Sequential discrimination process

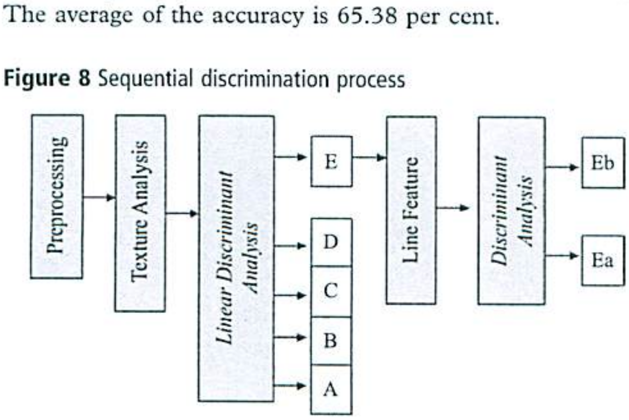


Table II Results of classification (after Step 1)

			Classification				
		Total	A	B	C	D	E
Manual classification	A	53	51	1	1	0	0
	B	71	3	52	15	0	1
	C	29	0	1	22	4	2
	D	41	0	0	2	31	8
	E	245	0	1	2	13	229
Accuracy (per cent)			94.44	94.55	52.38	64.58	95.42

Table III results of classification (after Step 2)

			Classification					
		Total	A	B	C	D	E_a	E_b
Manual classification	A	53	51	1	1	0	0	0
	B	71	3	52	15	0	0	1
	C	29	0	1	22	4	0	2
	D	41	0	0	2	31	4	4
	E_a	125	0	1	1	9	69	45
	E_b	120	0	0	1	4	53	62
Accuracy (per cent)			94.44	94.45	52.38	64.45	54.76	34.39

4. Conclusions

A sequential evaluation method combining two previously proposed classification scheme for protein crystallization was presented and evaluated with respect to manual classification by a human expert. The proposed method is an automated image-based scheme for classification of the protein crystallization state, involving pre-processing of images, first step classification into five Groups (A-E), and second step classification of Group E is into E_a and E_b types. The correct ratio of experimental result using the method presented here is approximately 70 per cent.

In future works, proposed method would be incorporated into TERA or automated crystallization systems as the crystal growth evaluation software and the performance evaluation of discrimination also would be done by the experiments.

References

Fisher, R.A. (1936), "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, Vol. 7, pp. 179-88.

Haralick, R.M., Shanmugam, K. and Dinstein, I. (1973), "Texture features for image classification", *IEEE Trans. Syst., Man, Cybern.*, Vol. SMC-3 No. 6, pp. 610-21.

Kawabata, K., Takahashi, M., Saitoh, K., Asama, H., Mishima, T., Sugahara, M. and Miyano, M. (2006), "Evaluation of crystalline objects in crystallizing protein droplets based on line-segment information in greyscale images", *Biological Crystallography, Actra Crystallographica*, Vol. D62, pp. 239-45.

Mahalanobis, P.C. (1936), "On the generalized distance in statistics", *Proceedings of the National Institute of Science of India*, Vol. 12, pp. 49-55.

Saitoh, K., Kawabata, K., Kunimitsu, S., Asama, H. and Mishima, T. (2005), "Evaluation of protein crystallization states based on texture information information derived from greyscale images", *Biological Crystallography, Acta Crystallographica Section D*, Vol. D61, pp. 873-80.

Sugahara, M., Nishio, K., Kobayashi, M., Hamada, K. and Miyano, M. (2002), "Development of the full-automatic protein crystallization and observation robot: TERA system", paper presented at International Conference on Structural Genomics (ICSG 2002), Berlin.

Corresponding author

Kuniaki Kawabata can be contacted at: kuniakik@riken.jp