Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Evaluation of crystalline objects in crystallizing protein droplets based on line-segment information in greyscale images

Kuniaki Kawabata,[a] Mutsunori Takahashi,[b] Kanako Saitoh,[a,b] Hajime Asama,[c] Taketoshi Mishima,[b] Mitsuaki Sugahara[a] and Masashi Miyano[a]*

[a]RIKEN (The Institute of Physical and Chemical Research), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan, [b]Saitama University, 255 Shimo-okubo, Sakura, Saitama, Saitama 338-0825, Japan, and [c]The University of Tokyo, 5-1-5 Kashiwa, Kashiwanoha, Chiba 277-8568, Japan

Correspondence e-mail: kuniakik@riken.jp

Received 8 May 2005
Accepted 8 December 2005

Several automated crystallization systems have recently been developed for high-throughput X-ray structure analysis. However, the evaluation process for the growth state of crystallizing protein droplets has not yet been completely automated. This paper proposes a new evaluation method for crystalline objects in automated crystallization experiments. The main objective is to determine whether a droplet contains crystals suitable for diffraction experiments and analysis. The evaluation method developed here involves extracting line-segment features from an image of the droplet and discriminating the state of crystallization using classifiers based on line features. In order to verify the efficacy of the proposed method, it was used to classify images obtained by an automated crystallization system.

1. Introduction

The dynamic development of protein crystal structure-solution methods with the application of high-throughput nano-crystallization robots requires new fully automated methods or systems for assessment of the crystallization trials. However, at present the task of identification of samples containing crystals suitable for use in X-ray diffraction experiments from a large number of crystallization trials is still conducted manually in most laboratories. One semi-automatic protein crystallization and observation robot system, TERA, has been developed at the RIKEN Harima Institute (Sugahara & Miyano, 2002; Sugahara *et al.*, 2002). It incorporates a scoring system that evaluates the growth condition of a crystallization solution sample by matching it with score-decision criteria as shown in Fig. 1 and recording the score.

The scoring operation is conducted by an expert, who visually inspects images of crystallization-droplet samples automatically recorded by a robot. If the droplet contains a crystal suitable for X-ray diffraction analysis, a score between 6 and 9 is assigned. In other cases, a score between 0 and 5 is assigned. Despite the capacity of TERA to obtain 500 000 images per month, the number of images that can be manually processed is limited. Therefore, there is a strong need to automate the score-decision process.

Previous studies on automated growth evaluation of crystallization solutions employed various methods, including the utilization of polarized filters (Bodenstaff *et al.*, 2002), a rotating polarizing filter (Echalier *et al.*, 2004) and image processing (Cumbaa *et al.*, 2003; Zuk & Ward, 1991). Rupp (2003) used phase congruency to detect a large number of small crystals and Gester *et al.* (2003) automated the counting of the number of crystals to generate three-dimensional surface plots of the crystals and to determine the crystal size based on the length of the perimeter of the crystals. Miyatake

et al. (2005) developed an automated crystallization/observation robotic system, HTS-80, which was reported to be able to categorize the crystallization droplet status into four stages based on the extracted contour information. Most of these previous studies focused on the existence or absence of a crystal. No detailed study to determine whether the detected crystals are suitable for X-ray diffraction experiments has been conducted thus far. Nonetheless, for efficient high-throughput protein structure analysis it is important to determine whether a specific crystal is suitable for X-ray diffraction analysis from a large number of crystallization droplets.

The present study aims to design and develop a method that automatically determines whether a crystallization droplet contains crystals suitable for X-ray crystallography. A previous study conducted by Saitoh *et al.* (2004, 2005) resulted in a highly accurate automatic determination of scores 0, 1, 2, 3 and 4–9. In this study, the images of crystallization solution with scores between 4 and 9 are classified into either category A, where the solution contains crystals suitable for X-ray diffraction analysis (scores between 6 and 9), or category B, where the solution did not contain crystals suitable for X-ray diffraction analysis (scores 4 and 5) (Fig. 2).

Table 1
Listing of size criteria for the protein crystals.

	Score 5 (category B)	Scores 6–9 (category A)
Dimension		
Length (longest side)	Less than 0.05 mm	0.05 mm or more
Thickness	Less than 0.01 mm	0.01 mm or more
Width	Less than 0.01 mm	0.01 mm or more

A score of 4 indicates the existence of amorphous grain which is not crystalline. Scores between 5 and 9 indicate the existence of protein crystals. A score of 5 (microcrystal) indicates that the sample cannot be considered for X-ray diffraction analysis owing to insufficient size of the crystals. The relationship between the scores and the crystal sizes is shown in Table 1.

Fig. 3 shows some typical examples of pictures taken using the automated TERA crystallization system. The method proposed in this study was evaluated using images which contain several growth-status crystalline objects (amorphous grain, microcrystal and crystal) in the same droplet. For example, some of the droplets contain amorphous grain and microcrystals and others contain microcrystals and crystals.

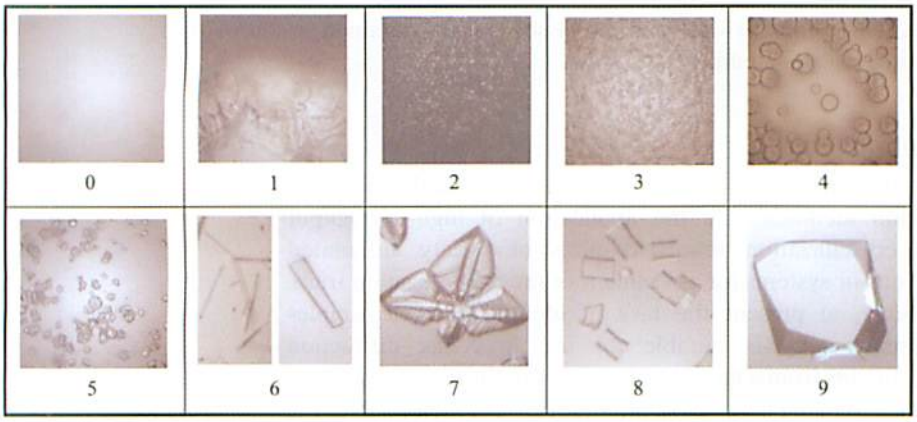

Figure 1
The ten categories for evaluation used by the RIKEN TERA system. 0, clear; 1, precipitate (i); 2, precipitate (ii); 3, precipitate (iii); 4, amorphous grain; 5, microcrystals; 6, needle- or plate-shaped crystals; 7, cluster of crystals; 8, crystal (i); 9, crystal (ii). (These images are a part of the complete droplet.)

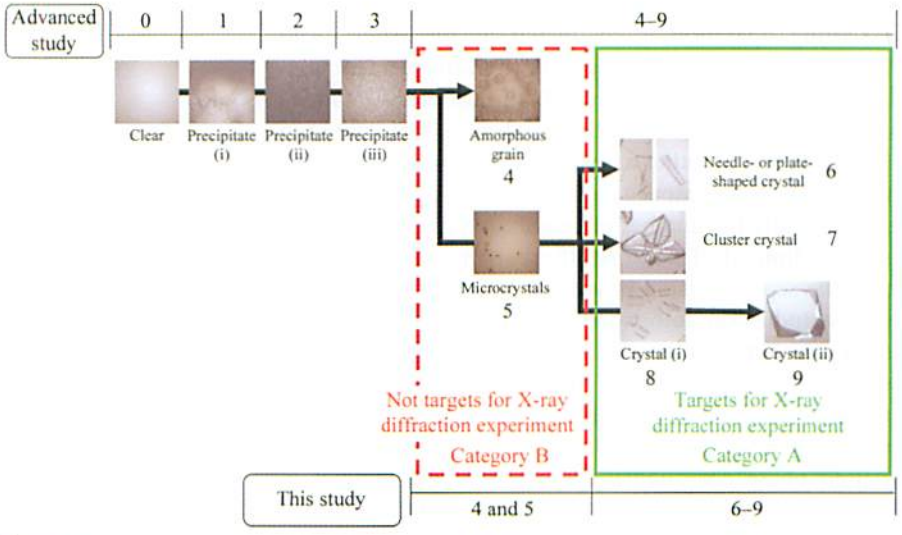

Figure 2
Evaluation categories and categorization into categories A (scores between 4 and 5) and B (scores between 6 and 9) in this study.

2. Proposed method

In the proposed method, a pattern-recognition process was used to determine whether the crystallization droplet contained crystals suitable for X-ray diffraction analysis. The images of the droplet are categorized into two classes: one in which the image contains crystals suitable for X-ray diffraction and one in which the image does not. The pattern-recognition process consists of a sequence of preprocessing, feature extraction and classification. An appropriate choice of visual feature values is important to ensure the accuracy of the determination process. In the manual process, during the second half of the growth period, droplets with scores between 4 and 9 are examined by an expert to determine whether the crystals are suitable for X-ray diffraction by

observing the shapes of crystalline objects (such as crystalline and non-crystalline grains) in the crystallization droplet. Thus, the characteristic shapes of the crystalline objects are regarded as characteristic contours and the lines which compose the shape contours of the crystalline objects are employed as feature values. Fig. 4 shows the processing flow of the proposed method.

2.1. Preprocessing

For the extraction of feature values, the images are subjected to preprocessing and the contours of the crystalline objects are extracted as binary edge images (Fig. 5). For this purpose, the original colour images are converted to 256-level greyscale images. Since the original images contain crystalline objects other than those in the crystallization droplet, such as the lateral side and the edge of the base part of the solution container, a part of the image containing the crystallization droplet is cropped from the greyscale images. The size of the cropped images is set at 450 × 450 pixels so that most of the base area is included. Next, an edge-detection process is conducted on the cropped images in order to determine the contours of crystalline objects. Of the various edge-detection methods, the Sobel filter operation (Takagi & Shimoda, 2004), one of the representative methods of edge detection in the image, is employed in this study. The Sobel filter operation followed by binarization processing is conducted on the edge-detected images to extract the contours of crystalline objects as binary edge images. For binarization, a certain threshold value is fixed and only those pixels with greyscales that are equal to or greater than this threshold value are considered to be the objects (or edge pixels).

However, this threshold configuration remains a key problem. Discriminant analysis (Fisher, 1936) is a threshold-setting method based on the density histogram. When crystallization droplet images are binarized by using the threshold value determined by discriminant analysis, several problems are encountered. In particular, a subtle concentration difference in the crystallization droplet is extracted as contour lines and the contours of the crystalline objects cannot be completely extracted. Therefore, in the proposed method, in order to specify a proper threshold value for binarization, the edge strength (caused by a subtle concentration difference of the crystallization droplet and the unevenness in lighting) is considered. 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) (Fig. 6) was investigated and it was found that a threshold value of 29 binarizes 99.9% of the pixels of the whole distribution as background. This threshold is therefore employed to extract the contour lines.

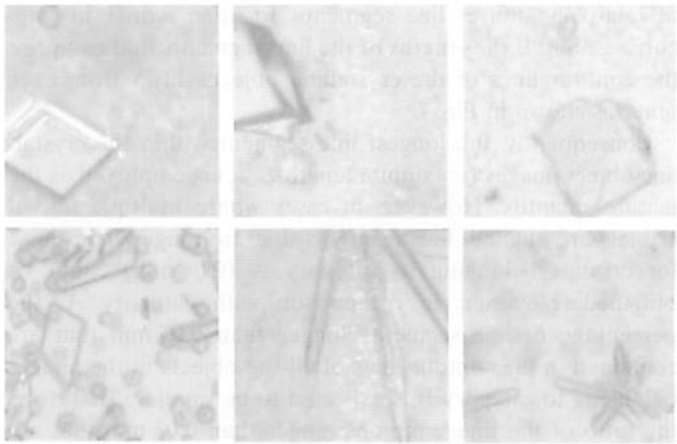

Figure 3
Examples of images obtained by TERA. Several growth states are seen to be present in the same droplet.

2.2. Extraction of feature values

Using the extracted contours from the binary edge images, the effective feature values for classification are extracted. When the contour lines of objects in categories A and B, with the scores shown in Table 1, are compared, it was found that the contour lines of objects belonging to category A are composed of longer line segments than those belonging to category B with a score of 5 (microcrystals). In addition, if the contour lines of the objects with a score of 4 (non-crystalline grains) are considered to be a collection of short line segments, the contour lines of crystals in category B must be composed

Figure 4
Processing flowchart of the method presented here. The method consists of preprocessing, feature extraction and classification.

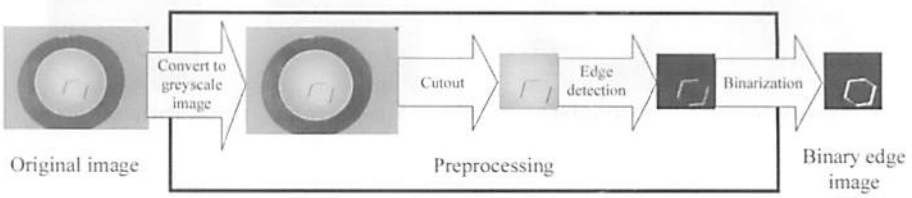

Figure 5
Preprocessing flowchart of the method presented here. The method consists of cutout, edge detection and binarization to obtain the binary edge image.

of relatively shorter line segments. In other words, in categories A and B the lengths of the line segments that comprise the contour lines of the crystalline objects differ from each other as shown in Fig. 7.

Consequently, the longest line segment within the crystalline object images (maximum length L_{max}) is employed as the feature quantity. However, in cases where multiple microcrystals are aligned, line segments that are longer than those for crystals belonging to category A (0.05 mm) could be obtained. However, in comparison with category A, the percentage of line segments longer than 0.05 mm that are contained in the contour lines of all the objects in the images belonging to category B is expected to be smaller. Therefore, the ratio of the linear regions longer than 0.05 mm that are contained in the contour lines of all the crystalline objects, R_{lsec}, is employed as a feature value. Furthermore, the number of crystalline objects in category B tends to be greater than for category A. Hence, the number of line segments in category B is also proportionally greater (Fig. 8). The number of line segments within the crystalline object images (N_{all}) is also employed as a feature value.

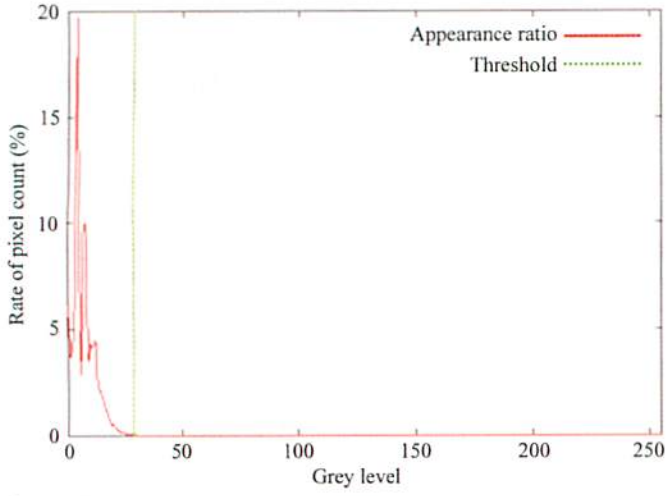

Figure 6
Edge-strength distribution of clear (score 0) images. The threshold value is 29, which binarizes 99.9% of the whole distribution.

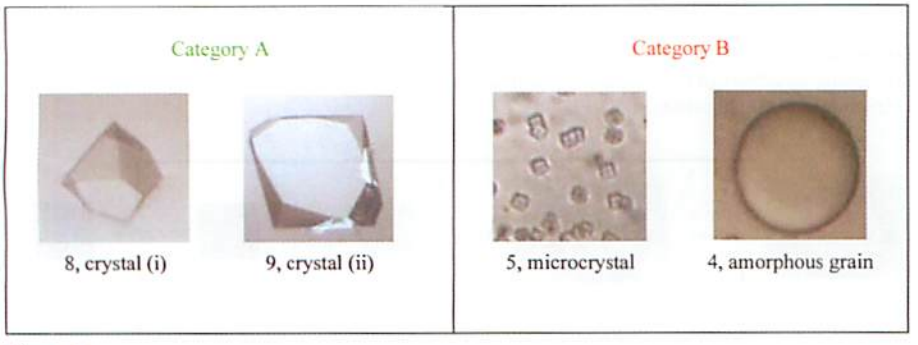

Figure 7
Contour lines representing the shape of the crystalline objects in the cutout images (100 × 100 pixels). The images in category A include a smaller number of longer line segments and the images in category B include a larger number of shorter line segments.

2.2.1. Extraction of the maximum length and the number of line segments. In order to extract L_{max}, the maximum length, and N_{all}, the number of line segments, from the feature values, 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 preprocessing, the crystalline object images need to be scanned from every possible direction. Thus, raster scanning is conducted by altering θ, the scanning direction ($0 < \theta < 180^\circ$), as shown in Fig. 9. An edge pixel that is detected on the scanning line when $n - 1$ line segments have already been detected is counted as the nth line segment and the pixel is considered to be the origin of the nth line segment $S_n(x_{sn}, y_{sn})$ and the scanning process is continued. The last edge pixel in a line segment is considered to be the terminus $E_n(x_{en}, y_{en})$ and the length of the nth line segment L_n is then determined as

$$L_n = [(x_{en} - x_{sn})^2 + (y_{en} - y_{sn})^2]^{1/2}. \quad (1)$$

This is added to the total number of line segments. For the detection of the total m line segments at the completion of scanning, the maximum length L_{max} and the number of line segments N_{all} can be evaluated, respectively, as

$$L_{max} = \max(L_n)\ (1 \le n \le m), \quad (2)$$

$$N_{all} = m. \quad (3)$$

2.2.2. Extraction of the ratio of linear regions. To extract the ratio of linear regions R_{lsec}, the number of linear regions longer than 0.05 mm within the contour lines of the crystalline object is determined. The extracted contour lines of the crystalline object could be distorted to some extent by noise during preprocessing. However, even if the original linear shape is not fully preserved, the approximate positional relation is sustained (Fig. 10).

As a result, the linearity of the region can be determined by observation of the positional relationship between the pixels within a region. In the case where all the pixels within a region are positioned in-line, three arbitrary points within a region are selected: two linear curves that connect the centre and two other points will essentially be at an angle of 180°. Moreover, if points adjacent to each other are not selected, even slightly distorted straight lines will form an angle similar to that made by the unaffected straight lines. Multiple sets of angles derived from two such lines connecting three points are investigated to evaluate the linearity of different regions. Firstly, contour tracking is performed on the binary edge images

and the extracted contour lines are sectioned by a width d. S^m indicates the mth region and S^{m-1} and S^{m+1} are the regions in front and to the rear of S^m, respectively. If the pixels belonging to each region can be defined as

$$S^{m-1} = (p_1^{m-1}, p_2^{m-1}, \ldots, p_d^{m-1})$$
$$S^m = (p_1^m, p_2^m, \ldots, p_d^m)$$
$$S^{m+1} = (p_1^{m+1}, p_2^{m+1}, \ldots, p_d^{m+1}),$$

then the angle between the line that extends from the kth point in S^m, p_k^m, to the kth point in the region S^{m-1} that is located at a width d backwards from p_k^m, p_k^{m-1}, and the line that extends from p_k^m to the kth point in the region S^{m+1} that is located at a width d forward from p_k^m, p_k^{m+1}, is defined as θ_{m_k}. The average angle in the region S^m is calculated as

$$\overline{\theta_m} = \frac{1}{d}\int_{k=0}^{d} \theta_{m_k}\,\mathrm{d}k. \qquad (4)$$

The three regions can be detected as one linear region since the three regions form a 180° angle if they are linearly positioned (Fig. 11). Here, an edge segment is considered to be linear if

$$170 \leq \overline{\theta_m} \leq 180° \qquad (5)$$

The contour lines of the whole image are evaluated and the ratio of the linear regions, R_{lsec}, can be evaluated as the ratio of the number of linear regions detected to the total number of regions. Since the detection of line segments that are longer than 0.05 mm (approximately 27 pixels in the image) is required and the three regions (width $3d$) are regarded as a single combined region when evaluating the contour lines, the width d is set at $d = 27$ pixels/3 = 9 pixels.

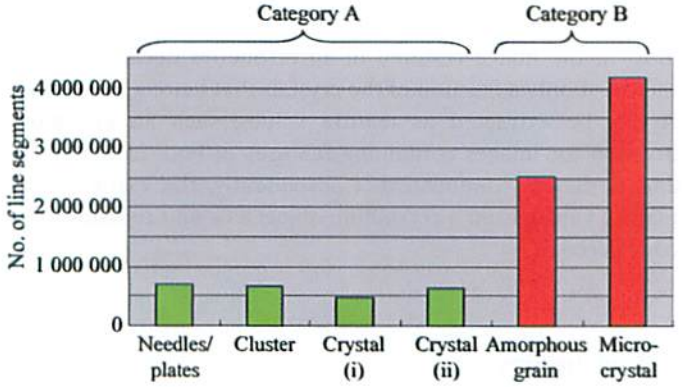

Figure 8
The number of line segments in categories A and B.

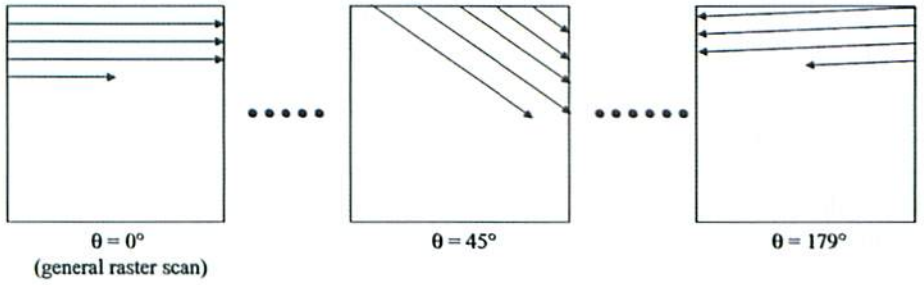

Figure 9
Raster scanning of the line segments on the cutout images in the direction θ. The detected line-segment information is utilized to evaluate the image.

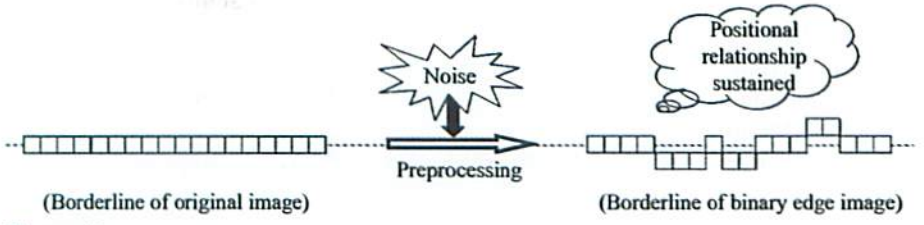

Figure 10
An example of the extracted line segment after preprocessing. Some of line segments in the original image become alternate line segments, with a zigzag shape, but sustain their positional relationship in the binary edge image.

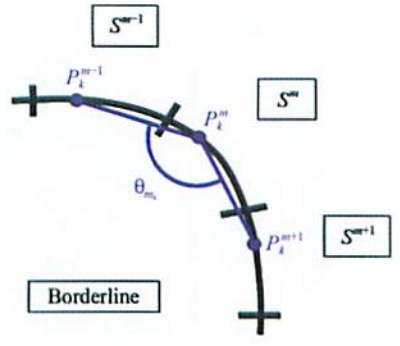

Figure 11
The evaluation of the contour lines of a whole image. Three connected line segments are utilized to evaluate the linear regions.

2.3. Determination using classifiers

$L_{\max}$, the maximum length of the extracted feature values, N_{all}, the number of line segments, and R_{lsec}, the ratio of the linear regions, are input into the classifier to determine whether the object images belong to category A or B. Various classifiers, for example self-organizing neural nets (Spraggon *et al.*, 2002), C5.0 (Bern *et al.*, 2004) and Bayes theorem (Wilson, 2002), have been used in previous studies for crystal image analysis.

We have investigated the various methods and have found that the method using discriminant analysis and Support Vector Machine (SVM; Vapnik, 1995), both of which are noteworthy for their efficiency in two-class identification, is the most efficient. Moreover, in the feature space consisting of the maximum length $L_{\max}$, the number of line segments N_{all} and the ratio of linear regions R_{lsec}, the variance between categories A and B differs significantly. Thus, Mahalanobis' generalized distance (Duda *et al.*, 2001) is employed as the classification criterion for the discrimination analysis. If the average of p variables in class k is $\mu_k = (\mu_1, \mu_2, \ldots, \mu_p)^T$ and

the inverse matrix of the variance–covariance matrix is $\sum_k^{-1}$, Mahalanobis' generalized distance of the observed value $X = (X_1, X_2, \ldots, X_p)^T$ and the class k can be evaluated by the equation

$$D_k^p = (X - \mu_k)^T \sum_k^{-1} (X - \mu_k). \quad (6)$$

3. Experiments and results

An experiment was conducted to verify the efficiency of the proposed method by using discriminant analysis either with Mahalanobis' generalized distance or with SVM as the classifier. Furthermore, the identification ability of discriminant analysis with Mahalanobis' generalized distance and SVM are compared. A library (available online at http://theoval.cmp.uea.ac.uk/svm/toolbox/) was used for SVM and the Gaussian kernel parameter σ and the parameter C, which adjust a soft margin, were assigned. For the experiment, we used 300 images that had previously been classified by a specialist (200 images from category A and 100 images from category B). The 300 images were preprocessed and the feature values extracted prior to classification by the different methods. The classification rates of the different methods were compared using the leave-one-out (l-o-o) method. The l-o-o method is one of representative evaluation and predicts the property value for a compound from the data set, which is in turn predicted from regression equation calculated from the data for all other compounds.

For the training set (300 images), a correct classification rate of 80.0% was achieved for the discriminant analysis and 88.7% for SVM (σ = 0.5 and C = 200, σ = 0.7 and C = 100). Furthermore, a 76.3% correct classification was achieved using the l-o-o method for the discriminant method and 88.7% in SVM (σ = 0.9 and C = 200), showing that SVM performs better than the discriminant method.

4. Discussion

When SVM is used for classification, the correct classification rate reached a value slightly below 90%, indicating the effectiveness of the proposed method. Currently, a fixed value is used for the binarization threshold, but improvements to the method could be made by the introduction of a threshold that would respond well to the variation in lighting conditions. This is a challenge to be considered for future studies.

In the feature space, the data distributions of categories A and B that are used showed a complex shape in the vicinity of the boundary. A higher capacity is shown by SVM than by discriminant analysis with Mahalanobis' generalized distance because a more complex boundary between distributions is possible with SVM. However, this needs further verification with a larger number of images.

The feature values that are employed in this study, the number of line segments and the ratio of the linear regions are evaluated for the whole image. This can result in incorrect classification when an image consists of both categories A and B by being affected by a region that contains a higher fraction of the image. Fig. 12 shows an image belonging to category A; however, characteristics indicating category B that occupy most of the image resulted in an erroneous decision. Nonetheless, if only a fraction of the crystals that belong to category A can be extracted as feature values, such an erroneous decision for images containing features of both categories A and B may be diminished. Consequently, the extraction of feature values from a crystalline object as a unit remains to be considered.

5. Conclusions

In this study, we aimed to improve the efficiency of protein crystal structure analysis and we proposed an automatic evaluation method for determination of crystal suitability for X-ray diffraction analysis. After contour lines of crystalline objects within an image are extracted, the feature values that are considered as effective for classification (the maximum length, the total number and the ratio of the linear region) are evaluated from the binary edge images. Two different classifiers, discriminant analysis with Mahalanobis' generalized distance and SVM, were applied to these feature values. The performance of each identifier is validated by the experiments.

Investigation of the feature-extraction method and the introduction of a threshold for binarization that is not affected by lighting conditions can be considered as future challenges. Furthermore, experiments should be conducted using more image data and the examination of classifiers other than the

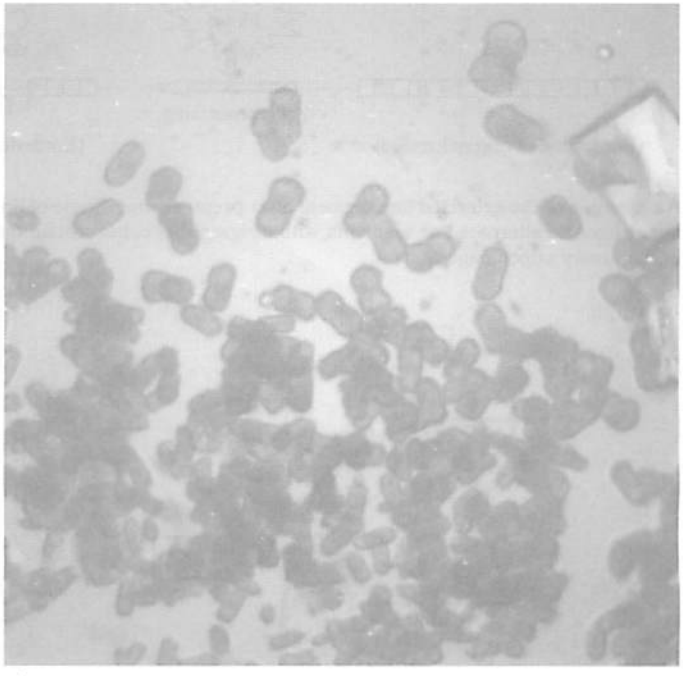

Figure 12
A typical example of an image containing crystalline objects belonging to both categories A and B.

discriminant analysis with Mahalanobis' generalized distance or SVM should also be addressed.

This method will be incorporated into TERA or other automated crystallization systems.

The authors thank Miss Maki Kumei, Mr Nobuo Okazaki, Mr Yuki Nakamura and Mr Tomoyuki Tanaka for supplying the protein sample images and for useful advice.

References

Bern, M., Goldberg, D., Stevens, R. C. & Kuhn, P. (2004). *J. Appl. Cryst.* **37**, 279–287.
Bodenstaff, E. R., Hoedemaeker, F. J., Kuil, M. E., de Vrind, H. P. M. & Abrahams, J. P. (2002). *Acta Cryst.* D**58**, 1901–1906.
Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., DeTitta, G. & Jurisica, I. (2003). *Acta Cryst.* D**59**, 1619–1627.
Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification.* New York: Wiley–Interscience.
Echalier, A., Glazer, R. L., Fülöp, V. & Geday, M. A. (2004). *Acta Cryst.* D**60**, 696–702.
Fisher, R. A. (1936). *Ann. Eugen.* **7**, 179–188.
Gester, T. E., Rosenblum, W. M., Christopher, G. K., Hamrick, D. T., DeLucas, L. J. &Tillotson, B. (2003). US Patent 6 529 612.
Miyatake, H., Kim, S.-H., Motegi, I., Matsuzaki, H., Kitahara, H., Higuchi, A. & Miki, K. (2005). *Acta Cryst.* D**61**, 658–663.
Rupp, B. (2003). *Acc. Chem. Res.* **36**, 173–181.
Saitoh, K., Kawabata, K., Asama, H., Mishima, M., Sugahara, M. & Miyano, M. (2005). *Acta Cryst.* D**61**, 873–880.
Saitoh, K., Kawabata, K., Kunimitsu, S., Asama, H. & Mishima, M. (2004). *Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems*, pp. 2725–2730. Piscataway, NJ, USA: IEEE.
Spraggon, G., Lesley, S. A., Kreusch, A. & Priestle, J. P. (2002). *Acta Cryst.* D**58**, 1915–1923.
Sugahara, M. & Miyano, M. (2002). *Tanpakushitsu Kakusan Koso*, **47**, 1026–1032.
Sugahara, M., Nishio, K., Kobayashi, M., Hamada, K. & Miyano, M. (2002). *ISGO International Conference on Structural Genomics, Berlin, Germany.*
Takagi, M. & Shimoda, H. (2004). *Handbook of Image Analysis: Revised Edition.* Tokyo: University of Tokyo Press.
Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Berlin: Springer-Verlag.
Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.
Zuk, W. M. & Ward, K. B. (1991). *J. Cryst. Growth*, **110**, 148–155.