Development of Crystallization Distinction Supporting System Using Image Processing

Kanako Saito[1], Kuniaki Kawabata[2], Satoshi Kunimitsu[2],
Hajime Asama[2], and Taketoshi Mishima[1]

[1] Department of Information & Computer Science,
Saitama University, 255 Shimo-okubo, Saitama, Saitama, 338-8570, Japan
Tel.+81-48-858-3723, Fax.+81-48-858-3723
kana@me.ics.saitama-u.ac.jp, mishima@me.ics.saitama-u.ac.jp

[2] Advanced Engineering Center, RIKEN,
2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
Tel.+81-48-467-9753, Fax.+81-48-462-4639
kuniakik@riken.go.jp, kunimitsu@ccl.riken.go.jp, asama@ccl.riken.go.jp

Abstract: In the post-genome era, it is one of important research subject to investigate the roles of the proteins in human body based on decoded genome information during Human Genome Project. In order to clarify them, it is necessary to analyze the structure of the protein crystals and their function.

Crystallization is the beginning stage of protein structure determination process. There are some methods for structural analysis of the proteins, and general one is X-ray structural analysis method. In order to utilize this method for analyzing the protein crystal's structure, artificial protein crystallization is required. However, since artificial crystallizing work takes much time and manpower, the performance against its cost is still low. Therefore, we started to discuss to develop a supporting system for improving efficiency of the crystallization distinction procedure. In this paper, we examine to realize such supporting system for crystallization distinction using image-processing technique and report about our experimental result with many real protein solution images.

1. Introduction

Genomic Sciences Center (GSC) of RIKEN (The Institute of Physical and Chemical Research) in Japan aims to determine the structures of 200-300 kinds of the protein crystals per a year in post-genome national project. In order to analyze the structures of the protein crystals, it is necessary to make the crystals artificially at beginning stage. Generally, every researcher always have 900-1000 protein solution samples for crystallization and keep to observe them until they are crystallized.

But now such crystallizing work is done by the researcher's handworking, and they don't have enough protein crystals against their needs. Therefore, in order to attain the project's aim, crystallizing work should be improved by automation technologies.

As crystallization techniques, there are Batch Crystallization, Liquid-Liquid Diffusion, Vapor Diffusion and Dialysis. Generally, Vapor Diffusion is the most popular one [3]. In Vapor Diffusion, there are two methods, hanging drop and sitting drop.

Hanging drop method (Fig 1): Using a tray with depressions, the protein solution is suspended as a drop from a glass cover slip above the precipitant solution in a sealed depression. The glass slip is siliconized to prevent spreading of the drop. Equilibrium is reached by diffusion of vapor from the drop to the precipitating solution or vice versa.

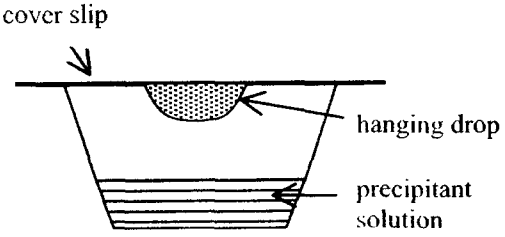


Fig 1. Hanging Drop Method

Sitting drop method (Fig 2): If the protein solution has a low surface tension, it tends to spread out over the cover slip in the hanging drop method. In such cases, the sitting drop method is preferable.

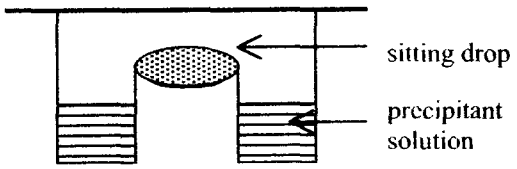


Fig 2. Sitting Drop Method

In this study, we utilize the images of the hanging drop diffusion style for experiments because the researchers of GSC use the style. In this paper, "the protein solution sample" indicates this sample.

Based on investigation result into the actual situation at the laboratory, we noticed that crystallization distinction procedure is not efficient.

Crystallization distinction procedure is basically to estimate crystallizing state in the protein solution. Crystallizing state of the protein is approximately classified into 4 groups (1:clear, 2:precipitate, 3:microcrystal, 4:crystal) as Fig.3 (based on the interview with the researchers of GSC).

Generally, it is unpredictable when the protein crystal appears in the drop. Therefore, it is always necessary to

observe all of the solution samples again and again. The researchers usually confirm all of the protein solution samples with their eyes through the microscope and save the image data to the computer by themselves. Such work takes too much time and high human-cost. In some case, they miss to find the crystals.

If such classifying work can be automated, the above-mentioned problems could decrease and such system contributes efficiency of crystallization distinction process.

Since conventional protein crystallization distinction depends on visual judgment through the microscope, we consider that the image information is important and useful. In this study, we investigate about the way of classification using image processing techniques and report experimental results of classifying the state of the protein solution sample.

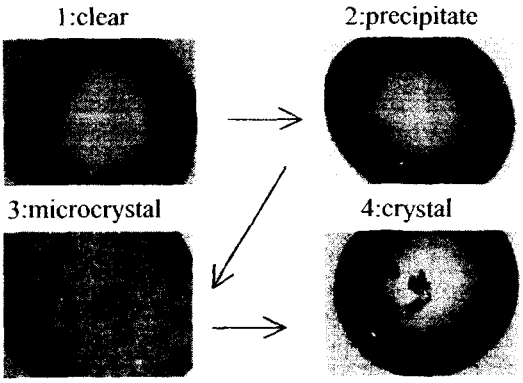


Fig 3. Approximate Categorization during Crystallizing in the protein solution

2. Features Extraction

2.1 Method of Feature Extraction

For crystallization distinction to the protein solution samples with image-processing techniques, we are utilizing static image data, which were photographed during crystallizing process at GSC. In this paper, we examine our distinction method using these image data. These data were evaluated by the researchers of GSC, beforehand.

Because how the protein solution samples transfigure is generally unpredictable, pattern matching or color distinction approaches can't be introduced in this case. Generally, each pixel's value in the image data is not independent to nearby pixel's value but there is any relation to each other. Usually, captured thing and visibility of any pixel are reflected on the nearest pixel, too. Nearest pixel is also arranged spatially with having close relation to the original pixel. Then, we consider that there is peculiarity pixel arrangement of each status of the protein solution sample. We also try to quantify such arrangements and extract the features.

Generally, "texture analysis", "smooth", "sharpen" and "edge detect/enhance" are utilized well for the quantification of the image. In the same distinguished category by the researcher, it seems that each image has similar characteristics. In this paper, we utilize texture analysis method to process the image of the protein solution sample.

2.2 Texture Analysis

Texture analysis method is a method for quantifying transition of the image tone. This method realizes to analyze two dimensional transition patterns of grayscale or color image, and it extracts important features of such image.

Texture feature values are usually calculated by using the co-occurrence matrix. Fig 4 shows basic derivation way of the matrix. The element of the co-occurrence matrix express the frequency of that the gray-scale value of point B that is distance $\delta = (r,\theta)$ (r: distance, θ : angle) from point A is j when the gray-scale value of point A is i. We derive fourteen sorts of feature values defined on Table 1 using the co-occurrence matrix, and characterize the pattern in the image with these values [2].

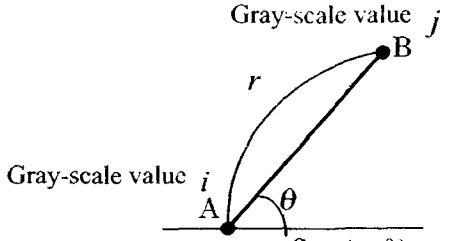


Fig 4. displacement $\delta = (r,\theta)$

Table 1. defined 14 features

1: angular second moment	9: entropy
2: contrast	10: difference variance
3: correlation	11: difference entropy
4: sum of square: variance	12,13: information measure of correlation
5: inverse difference moment	14: maximal correlation coefficient
6: sum average	
7:sum variance	
8: sum entropy	

2.3 Pre-processing

The original images of the protein solution samples are grayscale images that are photographed using the microscope of 40 magnifications, and their size is 1712[pixel] x 1368[pixel]. Before deriving the co-occurrence matrix of each image, we should transform the original images into appropriate size and gray-levels for easy handling.

Because crystallizing reacts in the drop of the solution sample, there are invisible and uneven areas on the original images. If we utilize the image for derivation of the feature as it is, we can't extract accurate features. Then, we introduce differential processing to the original image for solving this problem.

Sobel filter is known as a first-order differential filter. Fig 5 shows two types of actual filter for differential calculation. By these horizontal and vertical differential filters being applied to the original images and being combined, it is realized to eliminate uneven area in the image of the protein solution.

−1	0	1
−2	0	2
−1	0	1

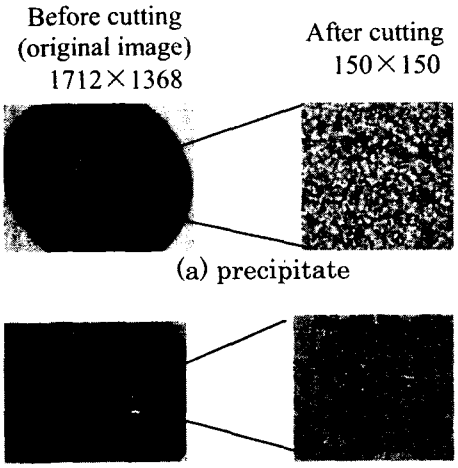
−1	−2	−1
0	0	0
1	2	1

(a) horizontal differential filter (b) vertical differential filter

Fig 5. Sobel Filter

As the above-mentioned conditions, the size of original images is 1712[pixel] x 1368[pixel] and it includes the whole image of the protein solution sample. However, it is hard to extract the characteristics from whole image of the protein solution. So we decided to divide the original images into small areas and picked up typical distinctive one. Here, since the size of microcrystal and crystal in the image is approximately 100[pixel], therefore we picked up 150[pixel] x 150[pixel] size area from each image (Fig 6).

Grayscale images generally have 256 levels and it takes high calculation load. Therefore, we transform 250 gray-level expression on the original image to 16 level expression as shown on Fig 7. Fig 8 shows pre-processing flow.



(a) precipitate

(b) microcrystal

Fig 6. Example of Sampled Images

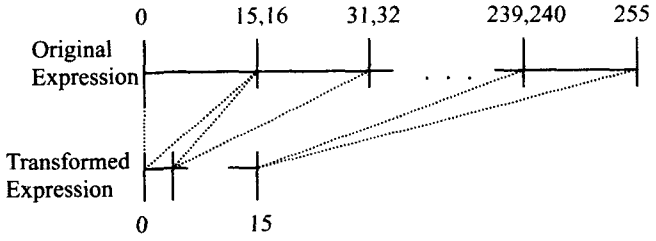


Fig 7. gray-level transformation

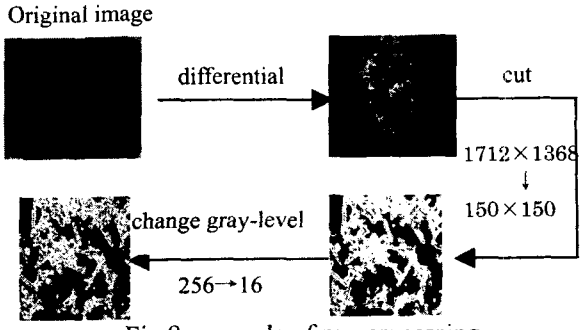


Fig 8. example of pre-processing

Using these pre-processed images, the co-occurrence matrix is derived. Here, we utilize $r=1$, $\theta=0°,45°,90°,135°$ as a set of the parameters $\delta=(r,\theta)$ for calculating the co-occurrence matrix (Fig 9). Calculated fourteen features are extracted from an original image by this procedure.

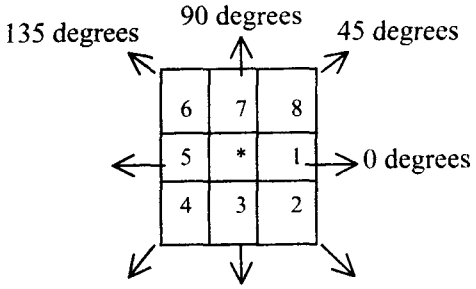


Fig 9. Pixel Arrangement Expression using θ (e.g. pixel 1 and 5 on 0 degree line and pixel 4 and 8 on 45 degree line)[2].

3. Distinction Process

3.1 Linear Discriminant Analysis

In this section, we try to classify the state of the protein solution sample into four categories (clear, precipitate, microcrystal and crystal) using extracted features using the co-occurrence matrix. Since each image has fourteen feature values, the dimension of feature space is fourteen, and it is hard to judge them by using each numerical value, intuitionally.

Multivariate analysis method is to distinguish a correlative multiplicity. There are some analyzing methods in multivariate analysis. Here, we utilize discriminant analysis method that is known as a standard technique of multivariate analysis. Discriminant analysis is to classify any sample data based on various features. And in this study, we utilize linear discriminant functions for distinction of the protein solution sample. Especially, we apply a linear discriminant method to distinguish the feature space into two groups. This is a technique to find suitable one-dimensional axis based on the pattern distribution in the feature space. Thus, such one-dimensional axis indicates linear discriminant function.

3.2 Discriminant Procedure

The discriminant procedure is as follows.

- By the LDA, the linear discriminant functions, which divide the sample data into the groups, is determined.
  g1: the function that divides the feature space into "clear" and "the other"
  g2: the function that divides the feature space into "precipitate" and "the other"
  g3: the function that divides the feature space into "microcrystal" and "the other"
  g4: the function that divides the feature space into "crystal" and "the other"
- Checking the correlation among feature values. If there is a pair having high correlation, then remove one of each other. Actually, we do not use 7:sumvariance and 14:maximal correlation coefficient.
- Calculation of discriminant scores of target image with the functions (g1, g2, g3 and g4).
- Finally, target image is classified into a group k : g_k, that takes maximum score (k =1-4).

The linear functions for distinction are as follows.

$$g_1 = 72.13x_1 + (-4.69)x_2 + 34.39x_3 + 0.24x_4 + 36.16x_5 + 1.00x_6 + 147.43x_7 + (-66.14)x_8 + 5.75x_9 + 26.63x_{10} + (-105.38)x_{11} + (-327.37)x_{12} - 71.49$$

$$g_2 = (-17.19)x_1 + 3.94x_2 + 4.13x_3 + 0.11x_4 + (-100.53)x_5 + 0.13x_6 + 2.35x_7 + (-14.16)x_8 + 1.31x_9 + (-37.47)x_{10} + 40.65x_{11} + 44.57x_{12} + 113.5844$$

$$g_3 = 22.96x_1 + (-4.02)x_2 + 9.61x_3 + (-0.43)x_4 + (-12.39)x_5 + (-0.04)x_6 + 29.00x_7 + (-12.85)x_8 + 6.37x_9 + 2.82x_{10} + 6.93x_{11} + (-18.68)x_{12} - 13.13$$

$$g_4 = (-23.35)x_1 + (-0.16)x_2 + (-37.14)x_3 + 0.51x_4 + 247.04x_5 + (-0.49)x_6 + (-100.16)x_7 + 75.00x_8 + (-16.82)x_9 + 74.31x_{10} + (-79.58)x_{11} + 19.27x_{12} - 232.34$$

Here, we used 60 images for calculation of the linear discriminant functions (clear: 18, precipitate: 28, microcrystal: 4, crystal: 10). We also used 211 image data (clear: 69, precipitate: 87, microcrystal: 32, crystal: 23) to examine the effect of our proposed method.

4. Experimental Result

In this section, we examine to compare calculated category using our proposed method (calculated category) with the category that is categorized by researcher (categorization by human) (Table 2).

As the result is shown on Table 2, the accuracy of each category are 'clear': 100%, 'precipitate': 79.3%, 'microcrystal': 43.7% and 'crystal': 73.9%, and the overall accuracy is 80.1%. According to this result, success rate to microcrystal is the lowest of the four categories. We consider that there are two reasons. One reason is that numbers of microcrystal image for experiment is so few. Therefore, small number of distinction misses effect to its accuracy, largely. Another one is that it is very difficult to define boundaries, between precipitate and microcrystal, and also between microcrystal and crystal. From that point, it is very hard to define typical microcrystal. Because there are microcrystals that look like precipitate, or crystal. Then, the crystallization distinction is almost done based on the researcher's empirical knowledge. In some cases, the researchers also distinguish the same image into different category.

However, the other categories keep enough recognition rates. It is useful level to support protein crystallization distinction.

As our future work, we try to categorize protein solutions into more detailed levels for corresponding to existing needs in the crystallization distinction work.

Table 2

		calculated category					
		CL	P	M	CR	total	result
Categorization by human	CL	69	0	0	0	69	100%
	P	0	69	18	0	87	79.30%
	M	0	11	14	7	32	43.70%
	CR	0	1	5	17	23	73.90%

(CL:clear, P:precipitate, M:microcrystal, CR:crystal)

5. Conclusion

In post-genome era, structure analysis of the proteins is one of important missions. Therefore, efficient analysis procedure is required in such scientific research field.

In this study, we attempt to develop a supporting system for protein crystallization distinction using image-processing technology.

For extracting feature values from the images, we utilized texture analysis using the co-occurrence matrix. We also classify into four categories (clear, precipitate, microcrystal and crystal) by using calculated feature values based on linear discriminant analysis.

In future work, we consider applying our method to many other images of protein solution. We also examine to utilize the other image-processing techniques for realizing high recognition rate.

References

[1] Mikio Takagi, Haruhisa Shimoda, "HANDBOOK OF IMAGE ANALYSIS", UNIVERSITY OF TOKYO PRESS, 1991.

[2] R.M.Haralick, K.Shanmugam and I.Dinstein, "Texture features for image classification", IEEE Trans.Syst., Man, Cybern., vol.SMC-3, no6, pp.610-621, 1973.

[3] Jan Drenth, "Principles of Protein X-ray Crystallography", Springer-Verlag New York, 1994.