

輪郭線情報に基づいた結晶化ドロップ内の析出物の状態判定

○高橋 六徳^[1], 川端 邦明^[2], 斉藤 佳奈子^{[1][2]}, 菅原 光明^[2],
浅間 一^[3], 三島 健稔^[1]

^[1]埼玉大学, ^[2]理化学研究所, ^[3]東京大学

An evaluation method for the image of crystalline objects in crystallizing utilizing contour information

○Mutsunori TAKAHASHI^[1], Kuniaki KAWABATA^[2], Kanako SAITOH^{[1][2]}, Mitsuaki SUGAHARA^[2],
Hajime ASAMA^[3], Taketoshi MISHIMA^[1]

^[1]Saitama University, ^[2]RIKEN, ^[3]The University of Tokyo

Abstract - In the process of protein structure analysis, growth state of protein crystallization droplet is evaluated by human experts. This study focuses on the function of evaluating whether the extracts in crystallization droplet are suitable for X-ray diffraction or not. Because human experts observe the shape and size of the extracts to evaluate the growth state of the extracts, we utilize to extract line feature from the image of crystallization droplet. This paper proposes new discrimination method which evaluates linearity with a high degree of accuracy and discusses its performance.

1. 緒言

X線構造解析によるタンパク質構造解析のハイスループット化を目指し、タンパク質結晶化作業の自動化が試みられている。しかし、結晶化ドロップの成長状態判定作業は人により行われているため、高効率化を妨げる要因となっている。結晶化ドロップの成長状態判定作業の自動化に関する研究には、偏光フィルタを用いたもの[1]等があるが、従来研究の多くは、結晶化ドロップ中の析出物が結晶か否かの判定は行っているものの、その析出物がX線回折の対象（一定以上の大きさまで成長した結晶）であるかについては議論していない。しかし、結晶化ドロップの成長状態判定の最大の目的は、X線回折の対象となるタンパク質結晶を発見することであるため、析出した結晶がX線回折の対象となるか否かという判定を行うことは重要である。

これまでに、理研播磨研究所で開発された自動結晶化観察ロボットシステム：TERAで撮影された画像を評価対象として用いた斉藤らの先行研究[2]によって、Fig.1の判定スコア基準における‘スコア0, 1, 2, 3, 4~9’の判定は高い精度で実現されている。そこで本研究では‘スコア4~9’のみを対象とし、これを前述のX線回折の対象となる析出物が存在する状態‘スコア6~9’：カテゴリA (CA)と、存在しない状態‘スコア4, 5’：カテゴリB (CB)へ判定することを目的とする (Fig.1)。

2. 判定対象の特徴と従来手法

本研究の判定の対象である‘スコア4~9’の成長状態については、専門家は析出物の形状を観察することで、それがX線回折の対象であるか否かを判断している。そこで本研究では、析出物の形状の特徴を輪郭線により評価可能と考え、輪郭線の特徴を判定に利用することにした。輪郭線の特徴としては、CAの析出物は主に長い線分で構成されていること、また、曲線を短い線分が連なったものとみなせば、CBの析出物は短い線分のみで構成されていると見る

Category of previous study	Score decision criteria	Category of this study
0	Out of the target of X-ray diffraction experiment	
1	0:clear	
2	1:precipitate(i)	
3	2:precipitate(ii)	4,5 Category B
	3:precipitate(iii)	
4~9	4:amorphous grain, 5:microcrystal, 6:needle & plate, 7:cluster, 8:crystal(i), 9:crystal(ii)	6~9 Category A

Fig. 1 Score decision criteria and target category

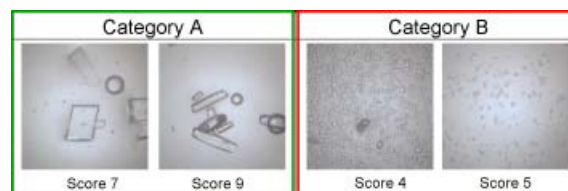


Fig. 2 Images of CA and CB

ことができる[3]。またタンパク質溶液が結晶化する際、はじめに多数の微少な結晶の核が出現した後、互いを吸収しながら大きく成長していくという過程をたどる傾向があるため、成長の後期にあたるCAの析出物の数がCBに比べて少なくなり易い傾向にあり、析出物の輪郭を構成する線分の数も少なくなる。よってCAの画像は小数の長い線分を、またCBの画像では多数の短い線分を含むという傾向を示す (Fig.2)。

以上に基づき、著者がこれまでに提案した手法 (以下、従来手法) [3]では、判定に利用する特徴量として、画像中存在する線分の本数 (線分数)、画像中でもっとも長い線分の長さ (最大長)、及び全体の輪郭に対する長い線分の部分の割合 (線分の割合)を用いた結果、約8割程度の確率で専門家と等しい判定を行うことが可能となっている。この従来手法において誤判定を起こした画像の大

半は、1)CB に属する画像で少数の大きなサイズの非晶質の粒が存在するもの (Fig. 3 (a)), 2)CA に属する画像で結晶以外に多数の沈殿物が存在する画像 (Fig. 3 (b)) のどちらかであった。

従来手法においては、析出物の輪郭線を現すエッジ画素が直線上に配置されている箇所を線分として検出していた。このため、実際の輪郭が直線となっていない箇所においても、Fig. 3 (a) のサイズの大きな非晶質の粒のように輪郭が緩やかなカーブを描く場合には、エッジ検出画像の輪郭線に幅があるため、エッジ画素が直線上に配置されることがある。また、析出物の輪郭線同士に重なりがある場合にも、同様にエッジ画素が直線上に配置されることがある。このため、大きな非晶質の粒が存在する画像や析出物の密度が高い画像は、X線回折対象となる結晶が存在しなくても長い線分が検出される場合があり、Fig. 3 のような画像の誤判定原因となっていた。以上のように従来手法では、長い線分に対する検出精度の低さが判定精度を低下させる原因となっており、検出精度を上げることで識別率を高めることが出来ると期待できる。

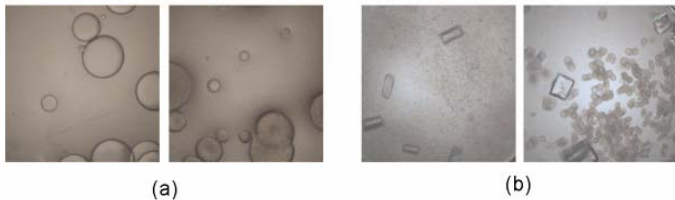


Fig. 3 Erroneous decision images on existing method

3. 提案手法

本稿では、輪郭線の直線評価のための新たな基準を追加し、高精度な判定を実現する手法を提案する。Fig. 4 に示したように、エッジの勾配の方向は、直線の輪郭線であれば一定の方向を向くが、直線以外の輪郭であれば方向にバラツキが出る。このバラツキを標準偏差で表すこととし、一定以上の長さを持ちかつ標準偏差の小さい線分の本数(長い線分数)を特徴量として追加した。またこれに伴い、提案手法における最大長の定義は、'長い線分と認められた線分の中でもっとも長い線分の長さ' とすることとした。以下提案手法の詳細について述べる。

まず、原画像に対し画像中の析出物の輪郭線を 2 値エッジ画像として抽出するための前処理を施す。続いて、結晶化ドロップ部のみが処理対象となるよう画像の切り出しを行う。さらに、切り出し画像に対して、グレースケール処理、Sobel 変換によるエッジ検出、2 値化を行うことで、2 値エッジ検出画像を得る。

特徴量である線分数、長い線分の数、最大長、を抽出するために、2 値エッジ画像に対して、走査方向 θ が $0 \leq \theta < 180[\text{deg}]$ のラスタ走査を行う。走査線上の連続したエッジ画素の集合を線分として検出し、本数カウントを行うことで“線分数”を得る。さらに、対象とした線分の長さが X 線回折対象の条件である $50\mu\text{m}$ (画像上では $27[\text{pixel}]$) 以上であったならば、エッジの勾配の方向の標準偏差を求める。この値が閾値以下であったならば、対象の線分を長い線分とみなして本数のカウントを行い、“長い線分数”を得る。また対象の線分長が、現在までに検出された長い線分よりも長い場合には、“最大長”を対象の線分の長さで更新する。なお本稿では、

複数の線分のサンプルから推定した確率密度関数を基に、エッジの勾配の方向の標準偏差の閾値を 0.11 とした。さらに直線区間の割合を抽出するために、析出物の輪郭線を一定の間隔に区切り、輪郭線全体に対してどの程度の割合で直線の区間が存在するのかを調べる。ある区間が直線であるかの評価するためには、区間内のお互いに十分離れた 3 点の作る 2 直線のなす角を複数の組について調べ、その平均値 $\bar{\theta}$ が $170 \leq \bar{\theta} \leq 180[\text{deg}]$ であったならば、この 3 区間を直線区間として検出する。対象画像全体の輪郭線を評価し、“直線区間の割合”を(検出された直線区間数)/(全区間数)として求める。

最後に抽出した特徴量を識別器に入力し、対象画像が CA・CB のどちらであるかの判定を行う。識別器には、2 クラス識別器として一般的である判別分析を用いることとした。また、特徴量空間における CA・CB の分散が多く異なることから、判別基準にはマハラノビス汎距離を用いた。

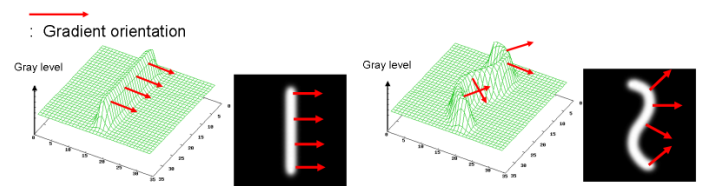


Fig. 4 Property of gradient orientation

4. 実験

以上の提案手法の判別有効性を検証するための実験を行った。実験には専門家による評価カテゴリが既知となっている 623 枚の画像データ (CA : 400 枚, CB : 223 枚) を用いた。また、比較のため従来手法を用いた場合の識別率も求めた。結果、学習データに対する識別率は、提案手法で 94.5%、従来手法で 87.5% となり、識別率が 7% 程度向上したことを確認した。

5. 結言

本研究では、結晶化作業の効率化のために、タンパク質結晶化ドロップが X 線回折実験対象を含むか否かを判定する課題に取り組んでいる。本稿では、以前提案した手法に改良を加えた手法の提案を行った。識別性能を確認する実験を行った結果、従来手法に比べて識別率が 7% 程度向上したことを確認した。

謝 辞

本研究は、財団法人大川情報通信基金の支援のもと行われました。この場を借りて厚く御礼申し上げます。

参 考 文 献

- [1] E.Rene Bodestaff et al, "The prospects of protein nanocrystallography", Acta Crst., D58, pp.1901-1906, 2002.
- [2] K.Saitoh et al., "Evaluation of Protein Crystallization States based on Texture Information" Proc. of 2004 IEEE/RSJ International Conference on IROS, pp.2725-2730, 2004.
- [3] K. Kawabata et al, "Evaluation of crystalline objects in crystallizing protein droplets based on line-segment information in grayscale images", Acta Crst., D62, pp. 239-247, 2006.