Published in final edited form as:

Cryst Growth Des. 2015 ; 15(11): 5254–5262. doi:10.1021/acs.cgd.5b00714.

CrystPro: Spatiotemporal Analysis of Protein Crystallization Images

Madhav Sigdel[†], Marc L. Pusey[‡], and Ramazan S. Aygun[¶]

University of Alabama in Huntsville, Huntsville, Alabama 35899, United States, iXpressGenes, Inc., 601 Genome Way, Huntsville, Alabama 35806, United States, and University of Alabama in Huntsville, Huntsville, Alabama 35899, United States

Madhav Sigdel: ms0023@uah.edu; Marc L. Pusey: marc.pusey@iXpressGenes.com; Ramazan S. Aygun: aygunr@uah.edu

Abstract

Thousands of experiments corresponding to different combinations of conditions are set up to determine the relevant conditions for successful protein crystallization. In recent years, high throughput robotic set-ups have been developed to automate the protein crystallization experiments, and imaging techniques are used to monitor the crystallization progress. Images are collected multiple times during the course of an experiment. Huge number of collected images make manual review of images tedious and discouraging. In this paper, utilizing *trace fluorescence labeling*, we describe an automated system called CrystPro for monitoring the protein crystal growth in crystallization trial images by analyzing the time sequence images. Given the sets of image sequences, the objective is to develop an efficient and reliable system to detect crystal growth changes such as new crystal formation and increase of crystal size. CrystPro consists of three major steps- identification of crystallization trials proper for spatio-temporal analysis, spatio-temporal analysis of identified trials, and crystal growth analysis. We evaluated the performance of our system on 3 crystallization image datasets (PCP-ILopt-11, PCP-ILopt-12, and PCP-ILopt-13) and compared our results with expert scores. Our results indicate a) 98.3% accuracy and .896 sensitivity on identification of trials for spatio-temporal analysis, b) 77.4% accuracy and .986 sensitivity of identifying crystal pairs with new crystal formation, and c) 85.8% accuracy and 0.667 sensitivity on crystal size increase detection. The results show that our method is reliable and efficient for tracking growth of crystals and determining useful image sequences for further review by the crystallographers.

Introduction

The success rate of crystallization can be as low as 2% depending on the protein and how crystallization trials are set up.¹ The crystallization depends on numerous factors such as protein purity, protein concentration, type of precipitant, crystallization techniques, etc.

Correspondence to: Madhav Sigdel, ms0023@uah.edu; Marc L. Pusey, marc.pusey@iXpressGenes.com; Ramazan S. Aygun, aygunr@uah.edu.

[†]University of Alabama in Huntsville

[‡]iXpressGenes, Inc.

[¶]University of Alabama in Huntsville

Finding the right combination of all these factors that would lead to crystallization is challenging since it requires setting up thousands of crystallization trials with different combination of conditions using techniques such as incomplete factorial design.² Moreover, protein crystallization is not an instantaneous process but rather a temporal process. The time required for the growth of crystals can take from several hours to many months. A crystallographer may thus be faced with hundreds of thousands of images of crystallization experiments to be reviewed. Due to the abundance of images collected, manual review of the images is impractical. To remove the burden of checking every trial image one by one by a group of experts and to automate the crystallization experiments, high-throughput robotic set-ups coupled with imaging techniques are used. Typically, the crystallization trials are prepared in well-plates with hundreds of wells and multiple droplet locations at each well corresponding to the different crystallization conditions. The robotic arrangement is made to scan the well-plates capturing images at each droplet location in each well.³ Most of these automated systems try to classify the crystallization trial images according to the crystallization state of the protein solution. Trials with no positive sign for successful crystallization may be discarded. On the other hand, the likely crystals are further reviewed either to optimize the conditions to prepare better crystals or to determine its suitability for use in X-ray crystallography.

A popular way of helping crystallographers using an automated system is to classify and score images collected by the robotic systems. The number of categories that are identified for crystal trial image classification may give different levels of information about the crystal growth process or optimization of conditions for further trials. At least trial images should be classified into crystal and non-crystal categories to identify experiments leading to successful crystallization (Zuk and Ward (1991),⁴ Cumba et al. (2003),⁵ Cumba et al. (2005),⁶ Zhu et al. (2004),⁷ Berry et al. (2006),⁸ Pan et al. (2006),⁹ Po and Laine (2008)¹⁰). To provide more information about the crystallization process, the protein crystal growth can be divided into multiple phases. Yang et al. (2006)¹¹ classified the trials into three categories (clear, precipitate, and crystal). In our previous work,¹² we classified images into non-crystals, likely leads, and crystal categories. Bern et al. (2004)¹³ and Saitoh et al. (2006)¹⁴ classified images into five categories (empty, clear, precipitate, microcrystal hit, and crystal). Likewise, Spraggon et al. (2002)¹⁵ and Cumba et al. (2010)¹⁶ have proposed six categories for the classification. In,¹⁷ we focused on the classification of crystal categories only. In these research studies, each image of a plate well or droplet location is analyzed individually without considering the previous collected images. For example, it is not possible to determine whether crystals are growing or the number of crystals is increasing without comparing with the history of images collected for that specific well or droplet location.

Besides the accuracy of classifying protein crystallization trial images, another major issue is effective extraction of features for developing real-time stand-alone systems. Systems that utilize high-performance, distributed, or cloud computing extracted as many features as possible hoping that the huge set of features could improve the accuracy of the classification.¹⁶ Some of the most commonly extracted image features are Hough transform features,¹⁰ gradient and geometry-related features^{10,17} blob texture features,¹⁴ features from differential image,¹⁴ Discrete Fourier Transform features,¹⁸ features from multiscale

Laplacian pyramid filters,¹⁹ histogram analysis features,⁶ Sobel-edge features,²⁰ gray-level co-occurrence matrix features,⁷ etc. While generating huge number of features may help to categorize images at a satisfactory level, it is not suitable for real-time systems where results should be delivered instantly as the plates are scanned. The image processing and feature extraction have been computationally expensive making it infeasible for real-time processing using stand-alone systems.

Analyzing temporal process of protein crystallization may alleviate the burden on the experts further. Recently, Mele et al. 2013³ described an image analysis program called Diviner that takes a time series of images from a crystallization trial and returns an estimate of the crystallographically relevant change between the images of the time course. The authors used the intensity change between images in a temporal image sequence as an indicator of crystallization (hopefully not dissolving of crystals). Diviner proposes the difference score for the overall image as a means to determine the importance of a crystallization trial. Their work leads researchers to work on images where a change in the crystallization droplet has been observed. However, the computational cost for the analysis is very high and not feasible for real-time analysis. Moreover, the pre-processing step of droplet (or drop boundary) detection proposed may not be effective for many crystallization trial images. For example, it may not be possible to detect the drop boundary because either the drop boundary is not present or the droplet is not circular.

Despite a few limitations, Mele et al.³ have shown that time series analysis may hint the experts about images to explore further. By analyzing the temporal images, we can visualize the changes in images and also derive some useful metrics regarding crystal growth. Information such as the appearance of new crystals, changes in the size of crystals and change in the orientation can be tracked from the temporal image sequences. In addition, time series analysis can also provide a relatively easier method for determining whether a temporal image sequence leads to crystal formation or not. In this study, our goal is to extract further information beyond the detection of change from time series images of protein crystallization trial images. The detection of change between images may not always carry useful image for the experts. Identifying whether 1) crystals are growing and 2) the number of crystals is increasing could be helpful information for crystallographers. Such spatio-temporal analysis of crystallization trials involves identification of regions, their growth, and matching them in the previous images. Since it depends on region segmentation and analysis of regions, there are further complications when time series images are analyzed due to the following reasons.

- i. Varying illumination, improper focusing, non-uniform shapes and varying orientation of crystals impose complexity in image analysis.
- ii. Since too many experiments are set up and the images are collected many times during the course of the experiment, image processing can take significant time. Existing automated systems have very high computational cost.

This work describes use of the CrystPro system for analysis of trace fluorescently labeled (TFL) protein crystallization results. Prior to setting up the crystallization trials, the proteins are trace fluorescently labeled (TFL) as described elsewhere.^{21,22,23} The basis of this

technique is to covalently label a low proportion, typically 0.2%, of the protein molecules with a fluorescent probe, then removing all probe molecules that are not attached to the protein. By removing all non-covalently bound probe, all fluorescence originates from a sub-population of the protein molecules. Thus the fluorescence is an indicator of the protein's behavior in the crystallization trial. When using the TFL approach, the fluorescence intensity is a function of the packing density of the covalently labeled protein. As the crystalline phase is the most densely packed, this leads to the principle that intensity equals structure. As the fluorescence is in the visible wavelengths, both automated and manual observations are feasible. As image analysis is dependent upon intensity and wavelength then this approach should be equally useful for images acquired using U.V. fluorescence.

EXPERIMENTAL SECTION

In this section, we provide information about material, experimental setup, and stages of our spatio-temporal analysis.

Material and Experimental Setup

The images of crystallization trials are collected using CrystalX2 software from iXpress-Genes, Inc. The proteins are trace covalently labeled with a fluorescent probe. As previously shown,^{21,22,23} the use of fluorescence significantly speeds the crystallization plate review process. The primary search criteria is intensity, not straight lines, which simplifies results interpretation by either software or direct visualization. Additionally, crystals that are obscured by other features in the well, such as buried in precipitate, growing in an out of focus location along the edge, etc. are readily apparent simply by the presence of their fluorescence intensity.

In our experiments, we have used green light emitting diodes (LEDs) as the excitation source. Carboxyrhodamine is used as the covalent labeling. We performed spatio-temporal analysis on three datasets - PCP-ILOpt-11, PCP-ILOpt-12 and PCP-ILOpt-13. Each dataset is captured from a 96-well plate with 3 sub-wells scanned 3 times at different dates leading to 864 images per dataset. In our experiments, the time gap between the first and second captures for PCP-ILOpt-11, PCP-ILOpt-12, and PCP-ILOpt-13 are 2 days. The interval between the second and third captures for PCP-ILOpt-11, PCP-ILOpt-12 and PCP-ILOpt-13 are 3, 3, and 4 days, respectively. The size of the dataset is limited due to the number of crystalline outcomes that can have gradual growth in crystal sizes and new formation of crystals. In these experiments, the hyperthermophile-derived protein pcg (pyroglutamate amino peptidase) was being subjected to optimization trials using a series ionic liquids at 0.1M concentration. The trace fluorescence labeling approach was necessary as the ionic liquids also often crystallized, giving false positives under white light, but not fluorescence, illumination. Expert scores are obtained by a knowledgeable observer (MLP) first closely examining and then scoring each well under white light microscopy. The scores are then reviewed by reference to the latest fluorescent images, to eliminate salt crystals from the high scoring outcomes. Regions of high fluorescence intensity that are not reflected in the first pass white light score lead to a second scrutiny under white light to resolve the source

of the intensity. Thus for manual scoring, fluorescence is used as a feedback mechanism to maximize the correct scoring.

Our image processing routines are implemented in Matlab. The user interface for image visualization is developed using Windows Presentation Framework (WPF). Microsoft Visual Studio 2012 is used as the IDE. On a Windows 7 Intel Core i7 CPU @2.8 GHz system with 12 GB memory, it took around 40 minutes (or 2400 seconds) to process (extract features from) 2592 images (3*864 images per dataset) in the first stage. The average processing time per image is about 1s. Hence, the analysis for spatio-temporal analysis between a pair of images takes around 2s.

Method Overview

Time series analysis for checking the number of crystals and change in size of crystals requires comparison of images collected at different time instances for a specific well or a droplet location. Consider a sequence of protein crystallization trial images captured at n different time instances represented as $I = \{I_1, I_2, \dots, I_n\}$ where I_k represents the image collected at the k^{th} time instant and $1 \leq k \leq n$. In Fig. 1, crystallization trial images corresponding to 3 protein wells that are captured at 3 time instances are shown. In the first image sequence (Fig. 1a), all 3 images consist of cloudy precipitates without crystals. Hence, this sequence is not interesting for the crystallographers. In the second (Fig. 1b) and third sequences (Fig. 1c), crystals are formed in the later stages of the experiments. In these images, the increase in the number or growing size of crystals can provide important information for the crystallographers. Since protein crystallization is very rare, only few experiments lead to crystals. Therefore, it is important to first correctly identify such sequences. Once such sequences are identified, an image in the sequence can be compared with any other image in the sequence leading to a combination of $C(n, 2)$ comparison of images.

Fig. 2 shows the basic components of CrystPro system. CrystPro consists of three major steps: 1) identification of trials for spatio-temporal analysis, 2) spatio-temporal analysis of identified trials, and 3) crystal growth analysis. Spatio-temporal analysis is especially relevant and useful on crystallization trials having crystals. Therefore, we first identify all the trials that have crystals having fair size in any image of a crystallization trial. By doing so, we eliminate all the trials which do not have crystals or the crystals are very small where spatio-temporal analysis is hard to achieve. Next, spatio-temporal processing is done on the selected trials and spatio-temporal metrics such as change in intensity, foreground area, number of crystals, etc. are obtained. These metrics are used to identify crystal growth such as new crystal formation, crystal size increase, etc. and to perform further analysis.

Identifying Trials for Spatio-temporal Analysis

For the spatio-temporal analysis, we first want to select the crystallization trials. Image thresholding is often used in image analysis for separation of foreground regions from the background. Obtaining a good binary image is very critical in image analysis because any error in the binary image will propagate to further processing steps. For example, regions that belong to a crystal should not be cropped in the thresholding stage. If a crystal region is

partially detected, then the growth of a crystal may not be analyzed properly. Varying illumination, improper focusing, non-uniform shapes and varying orientation of crystals impose complexity in separating the crystals correctly. Moreover, images with skin formation should be identified in this phase and should not be evaluated for crystal growth in size or new crystal formation. Otsu's thresholding²⁴ is a very popular method for thresholding images. Likewise, Canny edge detection²⁵ has been widely used in the literature to detect shapes of objects. In this study, since we look into new crystal formation and growth of crystal size for spatio-temporal analysis, the size of crystals should be comparable and not sensitive to poor binarization of images. Therefore, the likely crystals are expected to have closed regions and within a certain size range for the spatio-temporal analysis. A single technique may not work properly for all images. To identify images with crystals, we propose a method that utilizes the results from both techniques for detecting trials with crystals. Instead of extracting large number of image features and applying a training based system, we propose this method as it is quick and effective.

Image thresholding—Segmenting an image into regions and then determining the important regions for further processing are important processes in image analysis. A thresholding algorithm typically classifies pixels into two classes: background pixels (pixels having intensity lower than a certain threshold) and the foreground pixels. Numerous image thresholding techniques have been proposed in the literature. Otsu's method²⁴ iterates through all possible threshold values and calculates a measure of spread of the pixel levels in foreground or background region. Otsu's image segmentation method selects an optimum threshold by maximizing the between-class variance in a grayscale image. Fig. 3(b) provides results using Otsu's threshold on 2 images. Here, the first image (I_1) does not have crystals but the second image (I_2) does. By eliminating the very large non-crystal regions (for example, region size above 2.5% of image size), we get the binary images shown in Fig. 3(c). These results show that non-crystal regions might appear as the foreground as an outcome of the thresholding method.

Canny edge detection—Canny edge detection algorithm²⁵ is one of the most reliable algorithms for edge detection. Our results show that for most cases, the shapes of crystals are kept intact in the resulting edge image. In this study, we are focusing on crystals regions that have closed components. With this assumption, we eliminate all the unclosed edges. Fig. 3(d) provides the Canny edge image for the original images in Fig. 3(a). Fig. 3(e) shows the result after eliminating the unclosed regions. The final image (Fig. 3(e)) for the first image is blank, which suggests that there are no crystals. The final image for the second image has likely crystal regions.

Merging results of thresholding and Canny edge—We propose a method that uses the results from Otsu's thresholding and Canny edge detection to identify trials with large crystals. For reliable detection of crystal growth analysis, the region of a crystal should have at least 10 pixels for 320×240 image resolution. The size can be matched to other resolutions proportionally. Fig. 4 shows the basic flow of the method. For every image, Otsu's thresholding image is applied first and very small and very large regions are eliminated. If the resulting binary image has foreground objects, Canny edge image is applied and likely

crystal regions are determined. An image is considered to have crystal if it has likely crystal regions from the results from Otsu's binary image as well as Canny edge image. A crystallization trial is considered to have crystals if at least one instance of the trial image is found to have crystal. This method provides a quick and accurate approach to identify trials suitable for spatio-temporal analysis. The experimental results are provided in Section 3.

Spatio-temporal analysis of protein crystal growth

Given a set of images of a protein well captured at different time instances, we are interested in analyzing the growth of protein crystals. This analysis requires matching crystals in the two images and identifying the changes. Information such as appearance of new crystals, dissolving of crystals, changes in the size of crystals, etc. gives useful information about the growth of crystals. Fig. 5 shows the overview of the spatio-temporal analysis system. The first stage of this analysis is registration of images that are collected at different time instances. Since the robotic microscope may not have the same exact position each time a well is captured, the images should be aligned to make proper analysis. For an image pair (I_p, I_q) , image I_q is aligned with respect to I_p and is represented as I_q^p . Next, binary images are obtained for the images I_p and I_q^p represented as B_p and B_q^p respectively. The binary images are then matched and spatio-temporal metrics related to crystal growth changes are extracted. These spatio-temporal metrics are used to predict the crystal growth changes. In CrystPro, in addition to comparing consecutive images, the first image in the sequence is also compared with the last image in the sequence. This helps to detect overall change in the solution. For n images in a sequence, this results in n comparisons of images in the sequence. The steps of spatio-temporal analysis of protein crystal growth is described next.

Identifying crystallographically important regions—In general, the protein crystallization images containing crystals have 4 different intensity regions: background region, droplet (the solution), high intensity regions around crystals and the crystal regions with the highest intensity. The background is the least illuminated region. The droplet region has higher illumination than the background. If crystals are present in an image, crystal regions will have the highest intensity. Also, the regions around the crystals have high intensity because of the presence of crystals. For spatio-temporal analysis of crystallization images, we are only concerned with the shape, size and growth of the crystals. Our main purpose is to separate crystal regions from the rest of pixels. In our experiments, a single threshold produced poor results for some images.

Otsu's method²⁴ has been extended to generate multiple thresholds and categorize multiple classes of pixels in an image. For example, given an image, k thresholds can be determined to differentiate the image pixels into $k + 1$ classes according to intensity. To separate pixels in an image into four classes, three thresholds need to be calculated. Therefore, we apply multilevel thresholding with three threshold levels and identify four intensity regions. The binary image is obtained by considering the pixels above the highest threshold as the desired foreground. The rest of the pixels are considered background. Furthermore, region segmentation is applied and very small and very large regions are eliminated.

Fig. 6 shows the result of applying multilevel Otsu's method with three threshold levels on a sequence of crystallization trial images. In Fig. 6(b), we can observe how the pixels are separated into 4 intensity regions. The red pixels have the highest intensity and represent the foreground whereas the blue, cyan and yellow pixels represent the low intensity regions. By considering all the red pixel regions as foreground, we obtain the binary images shown in Fig. 6(c). It should be noted that the white regions in the binary image (Fig. 6(c)) do not necessarily represent crystals. Generally, if an image does not have crystals, the partial illumination of the protein droplet yields large foreground region in the binary image. Those regions are discarded by applying region segmentation and filtering out the regions that are larger than a certain threshold (e.g., 2:5% of the image area). Fig. 6(d) shows the final binary images after filtering the regions based on the region size. These binary images are used for further analysis. For the majority, this method correctly identifies the crystals. However, if in an image there are multiple crystals with different illumination, we might miss the crystals having less illumination.

Image registration and alignment—Since images are collected at multiple time instances, and also because of the inaccuracy in the acquisition system, the images are not exactly aligned. Even though there may not be any change in the image sequence, the pixels may not appear at the same position in the consecutive images. To compare pixels of different images, intensity-based image registration²⁶ is applied. Very slight rotation is possible if the plate is not located exactly as it was positioned before. Rotation is ignored in the registration a) since neither the plate nor the microscope can rotate and b) rotation by positioning the plate is minor or can be covered by translations. The microscope lens can usually maintain its distance to the plate properly, so scaling parameters are also not considered. Hence, image registration only involves computation of translation parameters (t_x, t_y) for the next image in the sequence. Fig. 7(a,b) shows images of a protein well captured at two time instances. Fig. 7(c) shows the overlaying of image in Fig. 7(b) on top of image in Fig. 7(a) after alignment. The translation parameters are $t_x=-8$ (8 pixels to the left) and $t_y=9$ (9 pixels to the bottom) for this example. For proper alignment, the second image is shifted to the left and bottom with respect to the first image.

Determining crystal growth changes—Consider an image pair (I_o, I_n) from a time-series image trial where I_n is a new image collected later than an old image I_o . To compare the changes in crystal growth, we use the binary images B_o and B_n . B_n^o represents the binary image after B_n is aligned with respect to B_o . All the background pixels are represented by 0 and object pixels are represented by 1. We use the method from²⁷ for comparison of two regions. To find the common object pixels, new object pixels in B_n^o and the disappeared object pixels from B_o , we first compute a sum image $B_{o,n}$ given by equation 1:

$$B_{o,n} = B_o + 2B_n^o \quad (1)$$

As in,²⁷ the binary image B_n^o is multiplied by 2 and added with binary image B_o . The sum image $B_{o,n}$ consists of 4 different values - 0, 1, 2 and 3. The object pixels common in both the images have value 3. We refer them as matched pixels. All the object pixels in B_n but not in B_n^o have the value 2. Such pixels are added pixels. All the object pixels in B_o but not in B_n^o

have the value 1. These are considered removed pixels. Likewise, all the background pixels in both the images have the value 0. Using this information, we calculate the following statistics.

- The number of matched pixels (N_m) = $\sum_{i=1}^W \sum_{j=1}^H (B_{o,n} == 3)$
- The number of added pixels (N_a) = $\sum_{i=1}^W \sum_{j=1}^H (B_{o,n} == 2)$
- The number of removed pixels (N_r) = $\sum_{i=1}^W \sum_{j=1}^H (B_{o,n} == 1)$
- Size increment factor (I_f) to evaluate the growth of crystals:

$$I_f = \begin{cases} 0 & \text{if } N_m < \tau \text{ (}\tau \text{ is the minimum threshold to consider a match; default value is 10)} \\ N_a/N_m & \text{otherwise} \end{cases}$$

Fig. 8(a–b) provides an image pair I_1 and I_2 collected at two time instances. Image I_2 is aligned with respect to image I_1 . Fig. 8(c–d) shows the binary images B_1 and B_2 obtained using multi-level Otsu's thresholding method. Fig. 8(e) shows the regions after B_1 and B_2 are matched. Here, all the matched object pixels are shown in white, all the additional pixels in B_2^1 are shown in green. The object pixels in B_1 missing in B_2^1 are shown in red. Numerically, the count of matched pixels is 2232 and the number of added pixels is 1223. Likewise, the size increment factor is equal to 0.55. This provides an estimation for increase in the size of the crystals in the second image.

Spatio-temporal features—To analyze the crystallization growth changes, we perform spatio-temporal processing and extract features from image pairs. Spatio-temporal features are compared for a pair of images (I_o and I_n). B_o and B_n represent the corresponding binary images. In our system, the dominant feature for indication of forming crystals is the increase in the intensity. The change of intensity ($\mu_d^{o,n} = \mu_n - \mu_o$) on the overall image can be used to determine whether crystals are forming. Similarly, the percentage change in the size of foreground area is computed as $A_d^{o,n} = 100 * |A_n - A_o| / A_o$. Since, μ_d and A_d provide rough information about the overall change, the analysis can be improved based on the segmented regions. The change in the number of regions is computed as $N^{o,n} = N_n - N_o$. To get further information whether the size of the regions increased or decreased, we determine the number of matched pixels $N_m^{o,n}$, the number of pixels in additional regions ($N_a^{o,n}$), and disappeared regions ($N_r^{o,n}$). Table 1 provides the list of spatio-temporal features and their descriptions. These features are extracted for every pair of consecutive images and also for the first and last image in the sequence. If there are k images in the sequence, $7 * k$ features are generated. These features are useful in the image review process.

Fig. 9 provides a sample sequence of images for spatio-temporal analysis. Rather than providing a simple scenario having clear large crystals forming or growing, we provide this example with small crystals and possible errors of segmentation to show that the combination of our spatio-temporal features can provide useful information even for such cases. In Fig. 9, new crystals appear in the second image (I_2) and growth of crystals are

observed in the third image (I_3). In I_1 , there is no crystal but thresholding method identifies a foreground region (Fig. 9d) which has higher intensity than its surrounding. Images I_2 and I_3 have many small crystals that may or may not match (Fig. 9(e–f)). It is possible to see growth of some crystals in I_3 . Therefore, the pair of (I_1 , I_2) can be used to test new crystal formation, and the pair of (I_2 , I_3) can be used to test crystal size growth.

Table 2 provides the spatio-temporal features for 3 image pairs. While the average intensity difference (μ_d) is not significant, other measures such as the percentage change in foreground area (A_d), difference in number of regions (N), etc. provide important information about changes in the sequence. In I_2 , the number of regions is increased by 15 compared to I_1 . Similarly, there are 3 new regions in I_3 compared to I_2 . The change between I_1 and I_3 is more significant. The number of additional regions in I_3 compared to I_1 is 17. The increase in the foreground area indicates the presence of crystal growth.

From results in Table 2, we may infer that new regions are forming in I_2 and I_3 , and crystal growth in size is observed in I_3 . Such information is useful for experts to gain knowledge about the phases of crystallization process after further review of the corresponding images. It is important to extract such information in challenging sequences as shown above where thresholding method may make mistakes. In the next section, we use these features to determine rules for predicting crystal growth.

Determining crystal growth—Formation of new crystals or growing crystals lead to three types of changes: 1) the overall foreground area, 2) the number of crystals, and 3) the size of a matching region (hopefully a crystal). Therefore, we use this information to guide us in identifying the type of protein crystal growth. Table 3 provides a list of rules used to predict if there is a crystal growth in a pair of images. The first rule is used to identify new crystal formation. The rule indicates that there must be some change in the number of crystals forming or disappearing. The second rule is used to identify changes in the size of crystals. This has two conditions that should be both satisfied. The first part states that the foreground area should increase by at least 50% for clear crystal growth. Similarly, the second part states that the area of the matched regions should increase by at least 50%. Any image sequence in which one of the rules given in Table 3 is satisfied is considered to have crystal growth and thus important from the crystallographer's point of view.

RESULTS AND DISCUSSION

In this section, we provide experimental results of spatio-temporal analysis of experiments and discussion of our results.

Identifying candidate trials for crystal formation and growth analysis

The first stage of new crystal detection or crystal growth in size is the detection of crystal regions in images. The classification results of identifying candidate trial images for 3 datasets are provided in Table 4. Each dataset has 864 images collected from 288-well plate collected at 3 time instances. The predicted results from our method are compared against expert scores. In Table 4, true negative (TN), false positive (FP), false negative (FN), and true positive (TP) refer to the number of trials correctly predicted as non-crystals, the

number of non-crystal trials predicted as crystals, the number of trials incorrectly predicted as non-crystals, and the number of trials correctly identified as crystals, respectively.

For PCP-ILOpt-11, the system predicts 24 (2 FP + 22 TP) trials with crystals. Similarly, for PCP-ILOpt-12, the system predicts 14 trials (6 FP + 8 TP) with crystals. Likewise, PCP-ILOpt-13 has 14 (3 FP + 11 TP) crystallization trials predicted to have crystals. The accuracy of the system is above 98% for all 3 datasets. Likewise, the sensitivity for detecting trials with crystals is also very high. There are few false negatives (missing trials with crystals). Those trial images are either blurred or have illumination problem. The system is able to reject non-crystals trials with very high accuracy. As successful trials are rare, this will help eliminate large proportion of unsuccessful trials.

Detection of new crystals

Formation of new crystals is an important outcome in crystallization experiments. Once our system detects formation of crystals and growing crystals, a crystallographer may start from these cases for his or her analysis. In Fig. 10, each column shows sample image pairs of crystallization trial images. In the image pairs shown in Fig. 10(a–b), no additional crystals are formed in the second image in comparison to the first image. On the other hand, Fig. 10(c–d) shows the image pairs where new crystals formed in the second image. Our CrystPro can distinguish these two scenarios and also provide the count of new crystals appearing in the crystallization trials.

For the crystallization trials identified with crystals from Section, temporal analysis for new crystal formation is carried out using *Rule 1* in Table 3. Let I_1 , I_2 , and I_3 correspond to images collected for a crystallization experiment at 3 different time instances. For each dataset (PCP-ILOpt-11, PCP-ILOpt-12 and PCP-ILOpt-13), the temporal analysis for 3 image pairs (I_1 , I_2), (I_2 , I_3), and (I_1 , I_3) are provided in Table 5. The objective is to determine the performance of the system to correctly identify new crystal formation by comparing the predicted outcome with the actual expert score. Since this step is based on the candidate trial identification for analysis, the candidate trials include both the true positive and the false positive trials. The PCP-ILOpt-11 and PCP-ILOpt-12 datasets did not yield any false negatives. The results shows that the system exhibits very high sensitivity for new crystal formation. Out of 78 total cases where new crystals are formed, 77 are identified correctly leading to satisfactory accuracy (performance) of our system.

Detection of crystal size increase

Analysis of the changes in size of crystals is another important factor for crystallographers. After identifying candidate crystallization trials for analysis, a spatio-temporal analysis between image pairs is carried out for detecting increase in size of crystals. Fig. 11(a–b) shows sample image pairs with no growth in the size of crystals. On the other hand, Fig. 10(c–d) shows the image pairs where crystals grow in size in the second images. CrystPro can detect crystal growth in size and distinguish these two cases using *Rule 2* in Section. The results for experiments for identifying increase in crystal size are provided in Table 6. The number of image pairs with observed crystal growth in size is 9, 5, and 4 for PCP-ILOpt-11, PCP-ILOpt-12, and PCP-ILOpt-13, respectively, considering pairs (I_1 , I_2), (I_2 , I_3), and (I_1 ,

I_3). CrystPro correctly identifies 7 out of 9 for the first dataset, 3 out of 5 for the second dataset, and 2 out of 4 for the third dataset.

Besides growth detection, CrystPro also determines other metrics such as the percentage of area change in the matched crystals. An expert may use these metrics for further analysis of the crystallization conditions.

Discussion

Trace Fluorescence Labeling

Influence of Labeling on Crystallization: The trace fluorescently labeled (TFL) approach requires the covalent attachment of a fluorescent probe to a subpopulation of the protein. We have shown that at the trace derivatization level ($< 1\%$) the presence of the probe does not affect the nucleation process, crystal growth, or crystal diffraction resolution.^{23,21} Other advantages that accrue from this approach are that the probe wavelengths can be selected to avoid interfering substances, one can use this method with direct visualization, and one can employ more than one color of fluorescent probe for the crystallization of complexes.

Possibility of Quenching: Protein crystallization screening involves trials over a wide range of components in chemical space. Some of these are likely to be fluorescence quenchers. While we have not observed strong quenching, it is highly likely that some may be taking place, manifested as reduced fluorescence intensity's. Quenching is a collisional process, with the degree of quenching being dependent upon the strength and concentration of the quenching species and the accessibility of the fluorescent probe. Fluorescent probes become shielded from the bulk solution environment upon incorporation into a crystal. Thus, while quenchers may be active on the probe while in solution they will be far less effective once the probes are buried in a crystal. Nonetheless, care should be taken when selecting a probe for TFL applications, particularly with respect to its pH sensitivity. In the experiments carried out to date we have not noticed any evidence of quenching.

Comparison with UV Fluorescence: Plate imaging systems primarily use visible light and UV fluorescence, which is dependent upon the protein having a tryptophan residue, which are not always present. Protein crystallization plate analysis using UV fluorescence requires UV transmissive optics, a UV light source, a suitably UV sensitive camera, and which for safety reasons cannot be used for direct visual observation. However, the UV fluorescence produces concentration dependent intensity, the same as for the TFL approach discussed here. Regardless, for both the UV and TFL methods, the signal of interest is the emitted intensity. Thus software methods developed for intensity-based image analysis and interpretation will be equally applicable with both UV and visible fluorescence. For this reason the CrystPro system should be equally suitable for UV fluorescence-based image analysis.

Spatio-Temporal Analysis

Increasing accuracy vs overfitting: Our system is not a supervised method, and hence is not affected by noise in expert scores. To improve the accuracy of results, one might be tempted to extract many features and train a classifier. However, it is very critical not to

overfit the data by using irrelevant features to build the model. Even careful selection of image features may not be sufficient to prevent overfitting for classification models. For example, to detect new crystals, the difference in the number of crystals is the only relevant feature. Likewise, to determine the growth of crystals, the changes in the size of matched crystals is a useful feature while the difference in crystal count is not. Table 7 provides the classification results using decision tree classifier with all the features in Table 2 as the input. Although this decision tree has good accuracy, the classification model is finely tuned by using irrelevant features (observed after analyzing the decision tree). Such a model may work on the training set but not on new crystallization trial images. The performance of our system mentioned in the previous sub-sections is a better indicator of the accuracy of our system. In addition, the performance of a system can increase as the number of simple image pairs increases or challenging image pairs are excluded. In our experiments, we provide true performance of CrystPro and rather included a stage of identifying candidate trials suitable for spatio-temporal analysis.

Size of crystals: CrystPro focuses on first identifying the crystallization trials that have crystals having some minimum size. Extremely small crystals are hard to match and susceptible to matching incorrect regions due to poor thresholding. Fig. 12(a) shows sample trial images where small sized crystals are grown but are missed by CrystPro. However, if the crystals grow in size, such images will be detected.

New crystal formation: CrystPro detects whether new crystals form or not. Hence, as soon as the plate is set up, the plate should be scanned and images should be captured before crystals grow. If the crystals are already grown in the first image, CrystPro does not detect new crystal formation. Fig. 12(b) shows sample trial image pair where crystals are fully grown in the first image and there is no apparent growth in crystals in the second image. To use CrystPro effectively, the images should be collected as soon as the plate is set up.

Proper thresholding: Having a proper binary image is very critical in automated image analysis systems. If the images in a temporal sequence are not thresholded properly, crystal regions may not be segmented properly. Hence, it cannot be determined whether those regions are growing or not. For some images, the proposed thresholding technique does not provide the correct binary image. Varying illumination and improper focus are some of the factors for incorrect thresholding. Normalizing the illumination is not always a good idea since crystal formation is indicated by the intensity increase in our experiments. If two images are normalized to have the same illumination, crystal formation or growth may not be detected. Therefore, effective thresholding is essential for the best results.

CONCLUSION

In this paper, we described our CrystPro system for automated analysis of protein crystallization trial images using time sequence images. Our approach involves generating proper binary images, applying image segmentation, matching the regions and then comparing the matched areas to determine the changes. Given the images of crystallization trials collected at different time instances as the input, we first identify the candidate trials for spatio-temporal analysis. Secondly, spatio-temporal analysis is done on the selected trials

and changes such as new crystal formation and crystal size increase are identified. This information is helpful for crystallographers to find out important conditions for crystal growth. The system has a very high accuracy and sensitivity for detecting trials with new crystal formation. Likewise, the system exhibits reasonable accuracy and high sensitivity for crystal growth detection. By analyzing the temporal images, we derive useful metrics to help the crystallographers review the images.

As future work, we plan to improve the image thresholding method. Normalization of images based on intensity causes problems in identifying growing crystals. If the images are normalized, the growing crystals may look like they shrink in time series analysis. We plan to use the background area for normalization rather than the complete image area. Such normalization may help to generate proper thresholded images.

Acknowledgments

This research was supported by National Institutes of Health (GM090453) grant.

References

1. Segelke BW. Journal of Crystal Growth. 2001; 232:553–562.
2. Luft JR, Newman J, Snell EH. Structural Biology and Crystallization Communications. 2014; 70:835–853.
3. Mele K, Lekamge BT, Fazio VJ, Newman J. Crystal Growth & Design. 2013
4. Zuk WM, Ward KB. Journal of crystal growth. 1991; 110:148–155.
5. Cumbaa CA, Lauricella A, Fehrman N, Veatch C, Collins R, Luft J, De-Titta G, Jurisica I. Acta Crystallographica Section D: Biological Crystallography. 2003; 59:1619–1627.
6. Cumbaa C, Jurisica I. Journal of structural and functional genomics. 2005; 6:195–202. [PubMed: 16211519]
7. Zhu X, Sun S, Bern M. Classification of protein crystallization imagery. Engineering in Medicine and Biology Society, 2004. 2004:1628–1631. IEMBS'04.
8. Berry IM, Dym O, Esnouf R, Harlos K, Meged R, Perrakis A, Sussman J, Walter T, Wilson J, Messerschmidt A. Acta Crystallographica Section D: Biological Crystallography. 2006; 62:1137–1149.
9. Pan S, Shavit G, Penas-Centeno M, Xu DH, Shapiro L, Ladner R, Riskin E, Hol W, Meldrum D. Acta Crystallographica Section D: Biological Crystallography. 2006; 62:271–279.
10. Po, MJ.; Laine, AF. Leveraging genetic algorithm and neural network in automated protein crystal recognition. Engineering in Medicine and Biology Society, 2008. EMBS 2008; 30th Annual International Conference of the IEEE; 2008. p. 1926–1929.
11. Yang, X.; Chen, W.; Zheng, YF.; Jiang, T. Intelligent Computing in Signal Processing and Pattern Recognition. Springer; 2006. p. 932–937.
12. Sigdel M, Pusey ML, Aygun RS. Crystal Growth & Design. 2013; 13:2728–2736. [PubMed: 24532991]
13. Bern M, Goldberg D, Stevens RC, Kuhn P. Journal of applied crystallography. 2004; 37:279–287.
14. Saitoh, K.; Kawabata, K.; Asama, H. Design of classifier to automate the evaluation of protein crystallization states. Robotics and Automation, 2006. ICRA 2006; Proceedings 2006 IEEE International Conference on; 2006. p. 1800–1805.
15. Spraggon G, Lesley SA, Kreusch A, Priestle JP. Acta Crystallographica Section D: Biological Crystallography. 2002; 58:1915–1923.
16. Cumbaa CA, Jurisica I. Journal of structural and functional genomics. 2010; 11:61–69. [PubMed: 20072819]

17. Sigdel, M.; Sigdel, MS.; Dinç, .; Dinç, S.; Aygün, RS.; Pusey, ML. In Emerging Trends in Image Processing, Computer Vision and Pattern Recognition. Morgan Kaufmann; 2015. p. 421-432.
18. Walker CG, Foadi J, Wilson J. Journal of Applied Crystallography. 2007; 40:418–426.
19. Xu, G.; Chiu, C.; Angelini, ED.; Laine, AF. An incremental and optimized learning method for the automatic classification of protein crystal images. 2006. p. 6526-6529.
20. Wilson J. Acta Crystallographica Section D: Biological Crystallography. 2002; 58:1907–1914.
21. Pusey M, Barcena J, Morris M, Singhal A, Yuan Q, Ng J. Acta Crystallographica Section F. 2015; 71:806–814.
22. Meyer A, Betzel C, Pusey M. Acta Crystallographica Section F: Structural Biology Communications. 2015; 71:121–131. [PubMed: 25664782]
23. Forsythe E, Achari A, Pusey ML. Acta Crystallographica Section D. 2006; 62:339–346.
24. Otsu N. Automatica. 1975; 11:23–27.
25. Canny J. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1986:679–698.
26. MATLAB, version 7.10.0 (R2013a). The MathWorks Inc; Natick, Massachusetts: 2013.
27. Dinç, I.; Dinç, S.; Sigdel, M.; Sigdel, MS.; Pusey, ML.; Aygün, RS. DT-Binarize: A Hybrid Binarization Method using Decision Tree for Protein Crystallization Images. 2014.

Synopsis

CrystPro is a spatio-temporal analysis system that is developed to detect the new formation of crystals and growth of crystals in size using time-series crystallization images benefiting from trace fluorescence labeling.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Sample temporal images of crystallization trials

Figure 2.
Overview of CrystPro system

Figure 3. Identifying images with crystals a) original images, b) Otsu thresholded images, c) eliminating very small and very large regions, d) Canny edge image, and e) regions with closed components in the edge image

Figure 4.
Process flow to identify crystallization images with crystals

Figure 5.
Overview of spatio-temporal analysis

Figure 6.

Image binarization using multi-level Otsu's thresholding, a) original image, b) image segmentation into 4 classes using Otsu's thresholding, c) binary image obtained by selecting the last class pixels as foreground and the rest as background, and d) final binary image after filtering minimum (5 pixels) and maximum (2.5% of image size) sized regions

Figure 7.
Spatial alignment using intensity based image registration a) Image at time instance t_1 , b)
Image at time instance t_2 , and c) Image 2 mapped on top of Image 1

Figure 8.

Matching images to determine crystal growth changes, a) original image I_1 , b) original image I_2 , c) binary image B_1 , d) binary image B_2 (aligned with respect to I_1), and e) matching B_1 and B_1^2 (matched object pixels in white, new object pixels in B_2 in green and removed object pixels shown in red)

Figure 9.
Sample image sequence with new crystals as well as crystal size growth

Figure 10.
Sample temporal trial images a–b) Image pairs with no new crystals, and c–d) Image pairs with new crystals

Figure 11.

Sample image sequences with increase in size of crystals in successive scan

Figure 12.
Sample crystal images with no distinct crystal growth

Table 1

Spatio-temporal analysis metrics

#	Metric	Description
1	$\mu_d^{o,n}$	Average intensity difference
2	$A_d^{o,n}$	Percentage of change in the foreground area
3	$N^{o,n}$	Difference in number of regions $ N_n - N_o $
4	$N_m^{o,n}$	No of matched object pixels
5	$N_a^{o,n}$	No of new object pixels
6	$N_r^{o,n}$	No of disappeared object pixels
7	$I_f^{o,n}$	Size increment factor

Table 2

Spatio-temporal metrics for image sequence in Fig. 9

Symbol	(I_1, I_2)	(I_2, I_3)	(I_1, I_3)
(t_x, t_y)	(-8,12)	(-7,11)	(-6,13)
μ_d	1.8	-0.6	1.2
A_d	105	63	240
N^-	15	3	17
N_m	35	824	52
N_a	770	655	1287
N_r	358	82	341
I_f	23	1.8	25.8

Table 3

Prediction rules for crystal growth

Rule #	Description	Rule signifying growth
1	New crystal formation	$N > 1$
2	Increase in crystal size	$A_d > 50\%$ AND $I_f > 50\%$

Table 4

Results of crystal detection on the 3 datasets

Dataset	TN	FP	FN	TP	ACC	SENS	PREC
PCP-ILopt-11	261	2	3	22	0.98	0.88	0.92
PCP-ILopt-12	273	6	1	8	0.98	0.89	0.57
PCP-ILopt-13	273	3	1	11	0.99	0.92	0.79

Table 5

Experimental results for new crystals detection

Dataset	Image pair	TN	FP	FN	TP	ACC	SENS
PCP-ILopt-11	I1, I2	7	6	0	11	0.75	1.00
	I2, I3	8	5	0	11	0.79	1.00
	I1, I3	5	7	0	12	0.71	1.00
PCP-ILopt-12	I1, I2	5	4	0	5	0.71	1.00
	I2, I3	7	3	0	4	0.79	1.00
	I1, I3	0	6	0	8	0.57	1.00
PCP-ILopt-13	I1, I2	3	2	0	9	0.86	1.00
	I2, I3	5	1	1	7	0.86	0.88
	I1, I3	3	1	0	10	0.93	1.00

Table 6

Experiments with crystal size increase

Dataset	Image pair	TN	FP	FN	TP	ACC	SENS
PCP-ILopt-11	I1, I2	20	2	0	2	0.92	1.00
	I2, I3	18	1	2	3	0.88	0.60
	I1, I3	20	2	0	2	0.92	1.00
PCP-ILopt-12	I1, I2	10	2	0	2	0.86	1.00
	I2, I3	8	5	0	1	0.64	1.00
	I1, I3	11	1	2	0	0.79	0.00
PCP-ILopt-13	I1, I2	12	1	1	0	0.86	0
	I2, I3	12	0	1	1	0.93	0.50
	I1, I3	12	1	0	1	0.93	1.00

Table 7

Classification results using decision tree classifier

Dataset	Test	Accuracy	Sensitivity
PCP-ILopt-11	New crystals	0.94	0.73
	Growth	0.97	0.83
PCP-ILopt-12	New crystals	0.89	0.90
	Growth	0.99	0.91