An Incremental and Optimized Learning Method for the Automatic Classification of Protein Crystal Images

George Xu, Casey Chiu, Elsa D. Angelini*, Andrew F. Laine
Department of Biomedical Engineering, Columbia University, NY
**Département Traitement du Signal et des Images (TSI), Paris, France*

Abstract

Protein production has experienced great advances in recent years. In particular, high throughput protein production, coupled with the use of robotics, outputs thousands of mixtures in micro-array wells. To detect the presence of protein crystal formation, images of these wells are acquired regularly using robotic cameras. Traditionally, a crystallographer would manually process each image – identifying the wells that resulted in protein crystal formation. This manual inspection process is slow and given the high rate of mixture output, it has become near impossible for crystallographers keep up. Our aim is to create an automated method of detecting which wells have crystals and which ones do not. We make use of a neural network that is trained based on manually classified ground truth data. After it is trained, the automatic classifier would give a binary output – a value of one for the detection of crystals and precipitates in images and a value of zero otherwise. In our previous papesr, the core methods of using multi-scale Laplacian image representation to extract image features and the implementation of the neural network classifier were discussed. Here we present a new, optimized approach to training the neural network and results from a large-scale test. We claim that the neural network can be better trained if the training image dataset is optimized in the sense that ambiguous images are removed during the initial training processes. Incremental training is implemented so that the network can be improved as more data becomes available. From initial results with training based on a 6,000 optimized image dataset, the accuracy of the improved classifier approaches 95% in identifying a wide array of images.

Introduction

The topic of genomics has garnered great interest in recent years. Numerous consortiums have been established with the goal of identifying and reproducing protein structures. The consortiums bring together scientists from a wide range of backgrounds. The process begins with biochemists who dream up the different formulations. These concoctions are then seeded using robotics with different mixtures on a 1565 matrix well plate to create a huge variety of cocktails. The mixtures are incubated and at regular intervals, such as a day, a week or two weeks, they are checked for protein formation using a robotic camera which acquires the image of each well. In the past, when the production output of concoctions was slow, a trained crystallographer would manually inspect each image to see which wells have resulted in protein crystal. Aside from crystals, the crystallographer may typically observe mixtures with precipitates, organic material, skins, no reaction, or a combination of all of these. Typically, protein formation occurs in 1% of all mixtures. The mixtures with crystals are further examined and once the protein structure is identified, the formulation of these mixtures is recorded and they can be placed in the protein production pipeline.

Given the recent advances and trend of moving towards high-throughput protein production, the identification of which mixtures resulted in protein crystals has become the bottleneck. Manual inspection of each well simply cannot keep up with the output. At the NESG Consortium, there are more than 3 million images that are backlogged for processing and this number grows everyday. This delays the discovery of concoctions that lead to protein formation and hence lowers the potential protein production output.

It is evident that an automated classification method is needed to address this problem. However, this problem is complicated by the fact that the protein structures may take many different shapes, precipitates and organic material not only clutter the image but also have similar shapes to protein crystals, the mixtures take the form of a droplet in the well which leads to irregular boundaries and non-uniform lighting conditions, and acquired images may be out of focus. Samples of these images are shown in Fig. 1. All this leads to difficulties in classification.

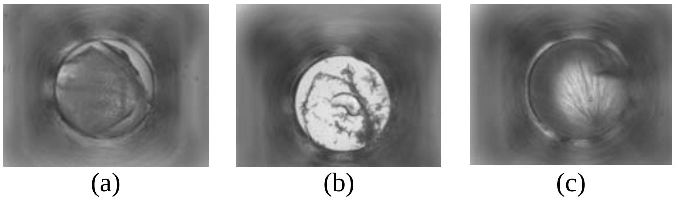
(a) (b) (c)

Figure 1: *Examples of well images. (a) & (b) both do not contain crystal or precipitates. (c) contains protein crystals*

The general approach we took to tackle this challenge is to use a neural network classifier. Previously expert labeled images are used as ground truth. Features from these images are extracted and pass in as inputs to the network. The classifier uses this previous knowledge to determine the likely presence of crystals in unknown images. The next section will provide an overview of the core methods and a detailed analysis for optimizing the training. The subsequent section will show the results of this effort. A detailed analysis of our core methods was presented in our previous papers [1,2].

Methods

Droplet cropping

The first step is to isolate the region of interest, which in this case is the droplet. The Ellipsoidal Hough transform is used to identify the possible elliptical boundaries in an image based on edge map information. The algorithm is modified as suggested by Malassiotis et al. in which gradient information is used instead of a computationally intensive three-dimensional search [3]. Figure 2a shows the result of performing this step.

Laplacian Pyramid Filter & Feature Extraction

The Laplacian filter is used to decompose the image into three different levels. Each level is obtained by subtracting a low-pass filtered image from the original image resulting in a pyramid structure shown in Figure 2. The Laplacian filter is used to extract the boundary information and image features. The multi-scale representation is capable of extracting the useful features of the image and at the same time, reduces the sample size. The image features are extracted from first and second order histograms of the Laplacian pyramid coefficients. The histograms contain useful information and clues as to the presence of protein crystals. Eight statistical features, which are invariant to orientation, are computed. These contain the mean, standard deviation, skewness, kurtosis, energy, entropy, autocorrelation, and power.

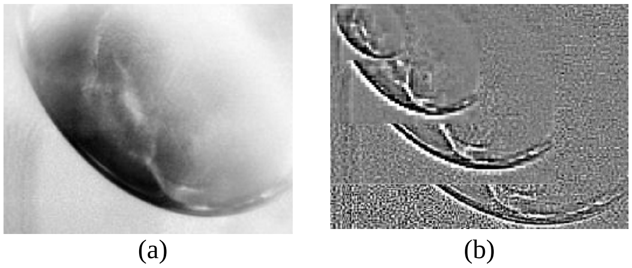
(a) (b)

Figure 2: *(a) shows the result from droplet cropping. (b) shows the 3 level Laplacian pyramid expansion*

Neural Network Classification

The feature vectors calculated in the previous step is passed into the input layer of a three-layer feed forward neural network shown in Figure 3. The LOG sigmoid transfer functions are used for both hidden and output layers to generate an output that is between zero and one. Backprojection and mean square error optimization were used to train the network. The output of this network is binary. An output of “1” would indicate the presence of crystals or precipitates as defined by the crystallographer who contributed the manually labeled image dataset used as the ground truth. An output of “0” would indicate that crystals or precipitates were not present but the image may contain organic matter, skins from a dried well, or a clear well. The actual value that the neural network outputs however is in between these two extremes. A threshold is needed to separate the two classes and this is discussed in optimization.

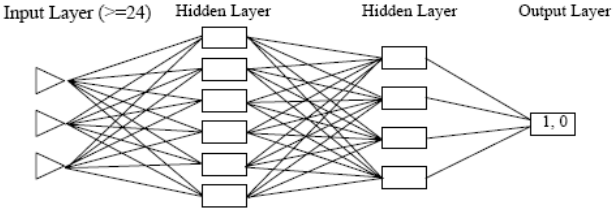

Figure 3: *Three layer feed forward neural network*

Training Database Optimization

The classifier is trained both using an optimized and random database of negative images. The optimized images are selected to minimize characteristics that tend to lead to ambiguous outputs, those which output from .4 to .6. To optimize the database, a sample dataset of roughly 200 images is run using a modified trained classifier. The modifications include capabilities to categorize problematic images. Images which outputted in the range of .4 and .6 are categorized under ambiguous and the falsely identified images are marked as either false positive or false negative. The images in the ambiguous category are manually scanned to determine possible universal characteristics or trends that led to their false classification or ambiguity. These characteristics can be broken into three broad categories: images with non-crystalline precipitates (Figure 4a), images with heavy ripples (Figure 4b), and images with significant air pockets (Figure 4c).

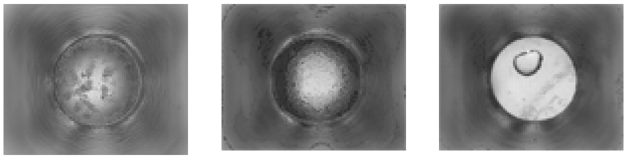
(a) (b) (c)

Figure 4: *Categories of ambiguous images. (a) non-crystalline precipitates (b) heavy ripples (c) air pockets*

The three image types are manually removed from a larger bank to produce sets of 1000, 2000, and 4000 “clean” images. The process to clean the dataset included viewing each image of the large dataset individually, and removing the image if it fell into the three categories. The reasoning behind the optimization is incremental learning. The classifier is initially trained with well-defined positive and negative images. Eventually, each training set fed to the classifier will contain an increased level of ambiguity.

Incremental Training

Given the large size and potential additions to the training database, it did not make sense to train the entire database at once. Instead a linear piecewise approach was taken and illustrated in Figure 5.

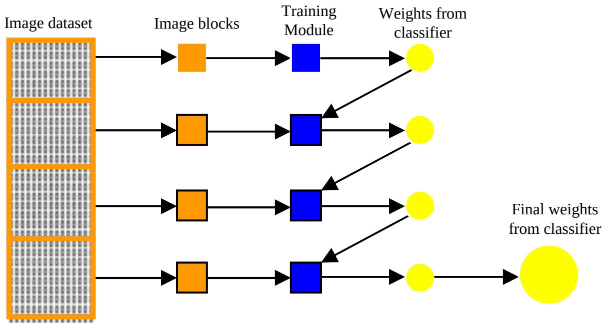

Figure 5: *Flow diagram of the training protocol. Calculated weights from a block of the image dataset are cascaded into the training of the subsequent image block*

From Figure 5, a block from the overall dataset is selected, used to train the neural network and the weights from the neural net classifier are saved. A second block of images is selected and this time, the weights from the previously trained neural net is passed as an input an additional input. The neural network uses the previous weights r to train with the new block of images leading to new weights for neural classifier which incorporates data from the previous two image blocks. This process is repeated until the training of the entire image database is complete.

An added benefit of this scheme is that as additional images are acquired, the weights of these images can be added to an existing network. This enables the neural net classifier to be updated without having to retrain the massive database.

Threshold Optimization

Given the rare occurrence of protein crystal formation (1% of all mixtures), it is crucial that an automatic classification method does not miss any hits. In other words, the false negative percentage needs to be kept as low as possible. Using some initial data, it was observed that a number of images with proteins did not surpass the default detection threshold of 0.5 and were labeled as a no protein/precipitate image. Figure 6 shows the classifier output for each image ranging from 0 for no crystals to 1 for crystal hits.

The solid dots are ideal or ground truth data and the circles are the classifier outputs. Ideally, the circles and dot should match. The red line represents the default threshold – everything above the line would be labeled as a crystal hit and everything below as a miss. It is evident that a number of images with crystals had a neural net classifier output that was below the red threshold line and falsely labeled as a miss. This problem was remedied by incorporating the blue dashed line which represents the new threshold line. This was experimentally determined and set at 0.03. Although the false positive percentages increase slightly, the more crucial statistic of false negatives fall drastically.

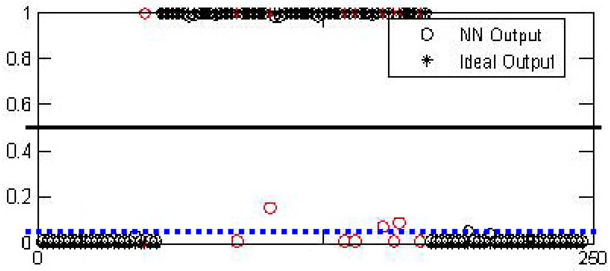

Figure 6: *Neural net classifier output. Red line is the original threshold, blue is the optimized threshold detection*

Results

In order to test the benefits of cleaning the training dataset, the classifier was trained using both uncleaned and cleaned datasets. The number of images in the database was also varied to find the effects of database size on the training of the classifier. These classifiers were tested by running a set of manually labeled, previously unseen images. The false positive percentage was obtained by finding the number of images without proteins or precipitates that were labeled as ones with it and vice versa for false negative. The total error was found by adding the false positives and negatives together. The ratio between images with proteins and precipitates and those without is roughly in a 1 to 10 ratio. While this ratio is greater than the actual percentages, it was found that this ratio yielded the best false negative rate. Using a smaller ratio tends to bias the classifier into generating more false negatives. Given the rare nature protein crystals, false negatives need to be kept at a minimum. The results are shown in Table 1.

Uncleaned Images

Size	**FP**	**FN**	**Error**
1,000	2.88	5.04	7.92
2,000	1.44	5.22	6.66
4,000	1.26	4.68	5.94
6,000	0.72	4.86	5.58

Cleaned Images

Size	**FP**	**FN**	**Error**
1,000	3.96	3.43	7.39
2,000	4.68	3.24	7.92
4,000	5.40	3.42	8.82
6,000	3.12	1.40	4.52

Table 1: *Results from testing*

From the results shown in Table 1, it is clear that classifiers trained with the cleaned image database yielded better results than those trained with the presence of ambiguous images. Furthermore, removing the ambiguity plays a larger role in small training datasets than compared to large datasets. With a database of over 4,000 images, the role of database optimization seems to have a much lesser effect presumably because the weight of a few ambiguous images is overshadowed by other more distinct images.

A second observation is that as the size of the training database expands, the classification results improve. The more images that the neural network is exposed to, the better is its abilities to classify unknown images. One way to improve accuracy is to add more manually classified images into the training. The largest training size that was used is 6,000 because there were not enough images with proteins to maintain the 1 in 10 ratio.

Discussion

The results from this classifier are very promising. While it may not attain the same accuracy as from manual inspection, the main advantage lies in that it can classify, with respectable results, completely automatically. While a trained crystallographer may be able to inspect 8,000 images a day, our automatic classifier can process 20,000 images a day without tiring. This number is scalable and dependent solely on computational power. It will continuously process images as long as the program and machines are running. While the ultimate goal is to replace the crystallographer and manual inspection all together, a semi-automatic method is still tremendously useful. Instead of processing 1565 images per well, the crystallographer can inspect only a handful that the automatic classifier had problems with and deemed as ambiguous.

Future work in this area aims to expand the training database to further enhance the reliability of the classifier. Since the current classifier only gives a binary output, another enhancement would be to increase the number of classification outcomes to include the ability to distinguish precipitates from crystals as well as organic matter and dried skins.

Acknowledgement

This project is part of the Northeast Structural Genomics Consortium (NESG) sponsored by the NIH for evaluating the feasibility, costs, economies of scale, and value of an infrastructure supporting high throughput structural genomics.

References

[1] E. D. Angelini, Y. Wang, A. F. Laine, "Classification of Micro Array Genomic Images with Laplacian Pyramidal Filters and Neural Networks", GENSIPS'04, Baltimore, MD, May 26, 2004.

[2] Y. Wang, D. H. Kim, E. D. Angelini, A. F. Laine, “Recognition of Micro-Array Protein Crystals Images using Multi-scale Representation,” SPIE International Symposium, Medical Imaging 2005, San Diego, CA, USA

[3] S. Malassiotis and M. G. Strintzis, "Tracking the left ventricle in echocardiographic images by learning heart dynamics," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 282-290, 1999.