The Automated Recognition of Protein Crystal Images by Feed-forward Neural Network

The production of protein crystals for x-ray crystallography is crucial for protein structure determination. Currently, it is difficult to predict the experimental conditions that will result in protein crystal formation. High-throughput experiments with varied crystallization parameters are often leveraged in the hopes that one or more parameters will lead to the formation of crystals. Images from multiple time points during micro-experiments are recorded and crystallographers are called upon to make a manual determination of crystal growth for each image. Consortiums in structural genomics such as NESG (North East Structural Genomics) now perform tens of millions of such micro-experiments annually resulting in the need to analyze an equally large number of images.

We describe a classification framework that is being developed in parallel with the NESG effort to assist and automate the screening for protein crystals. The three main components of the classification algorithm are multi-scale Laplacian pyramid filters, subsequent extraction of feature vectors used in a neural net classifier, and high-speed algorithm optimization. The processing steps include:

1. Field of interest cropping with a radii-weighed fast circular Hough transform.
2. Multi-scale image separation with Laplacian pyramidal filters.
3. Feature vector extraction from the histogram of multi-scale boundary images.
4. Feed forward binomial neural network classifier.

The feature vectors extracted from the histograms encapsulates geometric and textural features within each image and provides input to the neural network classifier. The feature vectors were themselves chosen based on PCA (Principle Component Analysis) of a number of potential discriminants.

A total of 79,632 images were independently classified by 3 crystallographers to be used for training and classification performance validation. Using a statistically weighed "hold-out" sample of 900 images, and using the crystallographers' ratings as ground truth, the current classification algorithm produced 88% true positive and 99% true negative rates (resulting in an average true performance of ~ 93.5%).

Recognition results are deposited directly into an Oracle database and made available through Macroscope data format. To enable NESG collaborators to efficiently perform crystal classification, incremental learning was implemented in the feed forward neural network. Similar to popular Bayesian email filters, users can sort particular images that were mis-classified into designated folders and the neural net will be automatically re-trained, thus enabling the classifier able to learn new crystalline structures.