AUTOMATED CLASSIFICATION OF CRYSTALLIZATION IMAGES

Christian A. Cumbaa and Igor Jurisica

Division of Signaling Biology, Ontario Cancer Institute, Toronto, Ontario, Canada

We report on attempts to use image analysis to automatically detect multiple, independent outcomes in microbatch-under-oil protein crystallization trials: phase separation, precipitate, skin, crystals, and clear drops. The study was based on a set of 147,456 images, generated and hand-scored for every outcome by Hauptman-Woodward Institute’s High Throughput Screening Lab.

Initial experiments at classification focused on training simple classifiers to recognize each outcome condition separately: one classifier to distinguish clear/unclear drops, one to detect presence/absence of phase separation, etc. Each image was fed to our image-analysis system, generating a set of 840 numeric features; these features were then fed to the classifiers specific to each potential outcome, which detect the presence or absence of each crystallization condition based on Gaussian models of the feature space.

The results from these initial experiments reveal differing degrees of difficulty in recognizing each outcome, with precipitate and clear drops being the most accurately detected outcomes (precipitate: 95% precision, 95% recall; clear: 85% precision, 97% recall), and crystals being the most difficult to detect (17% precision, 34% recall).

A review of these experiments, conducted jointly with our Hauptman-Woodward collaborators, highlights the need for improvement in three directions:

1. Better image features: no single image feature or pair of features from the set of 840 effectively separates the outcome-positive images from the outcome-negative images. Most features show heavily overlapping distributions of positive and negative images.
2. More powerful models: we must move beyond uni-modal distributions to successfully model the positive and negative images in feature space. Considering the variety of protein crystal morphologies (large crystals, microcrystals, needles), it is clear that the distribution of crystal-positive images will be multi-modal in any feature space.
3. Modeling of compound outcomes: we will present ongoing work in directly modeling compound outcomes, e.g., precipitate AND crystal AND NOT skin.

We have developed a new image analysis system that has revised and expanded the feature set generated from each image. This system focuses on the textural features of local regions of the image, offering more precise measures of texture, a finer-meshed screen for straight edge detection, and metrics on discrete objects found inside the well.

Image Analysis on the World Community Grid. With added features comes an increase in the computational costs of image analysis. We are in the process of converting our revised image-analysis system into a distributed system to be run on the World Community Grid. We intend to process the entire set of over $6 \times 10^7$ round-well microbatch images generated to date by HWI. This massive computing job is expected to start in the spring of 2007 and run for several weeks.