








Update May 2008

Thank you for your continuing support of the Help Conquer Cancer project. We are grateful for all the computing power you donate to this and other exciting and useful research at WCG. We do benefit

from it greatly, but we also participate in WCG as an Integrative Discovery Team:     . It is a TEAM effort (**T**ogether **E**veryone **A**ccomplishes **M**ore) that will help us to solve these complex problems.

Since the launch of Help Conquer Cancer project in November 2007, WCG members contributed almost 12,000 years of run time, averaging about 54 years a day.

Reminder about the complexity of protein crystallization

Crystallization is a multi-parametric process with three classical steps: nucleation, growth and cessation of growth. Technical difficulties in protein crystallization are due to mainly two reasons:

1. A large number of parameters affect the crystallization outcome, including purity of proteins, supersaturation, temperature, pH, time, ionic strength and purity of chemicals, volume and geometry of samples;
2. We only partially understand correlations between the variation of a parameter and the propensity for a given macromolecule to crystallize.

Conceptually, protein crystal growth can be divided into two phases: search and optimization. Search phase determines a subset of all possible crystallization conditions that yield promising crystallization outcome. These conditions are varied during the optimization phase to produce diffraction-quality crystals. Neither of the two phases is trivial to execute. If we consider only 20 possible conditions, each having 20 possible values, the result would be $1.04858E+26$ possible experiments; impossible to test exhaustively. Even a broad search phase may not produce any promising conditions, and many of the promising leads may elude optimization strategies.

High-throughput screening (HTS) can speed up the search phase, and has the potential to increase process quality. Automated image analysis and classification achieves two important goals: it improves throughput and generates consistent and objective results. Objective image classification is a necessary input to data mining and reasoning, which is essential to elucidate knowledge from large number of successful and failed crystallization experiments. These results will help understand protein chemistry and lead to achieving our overall goal – to improve number and quality of protein structures determined. We **hypothesize** that *(1) comprehensive and probabilistic image classification will increase both specificity and sensitivity of the process, and (2) systematic image analysis combined with data mining and reasoning will lead to improved understanding the chemistry of protein crystallization, and thus will also increase number of solved structures from the HTS pipeline.*



The challenge is the wide diversity of crystalsⁱ, as shown in Figure 1. To cope with this diversity, we must use multiple algorithms to identify crystals reliably, i.e., with high sensitivity and specificity.

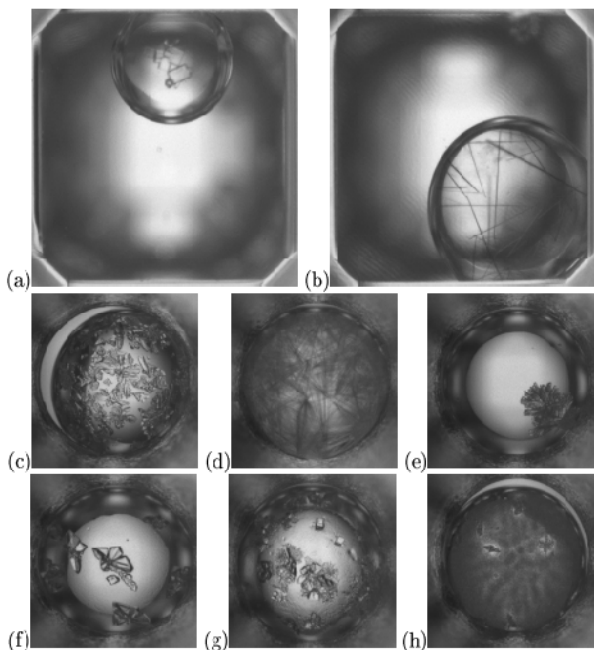


FIGURE 5.22: Examples of diversity by which crystal forms appear.

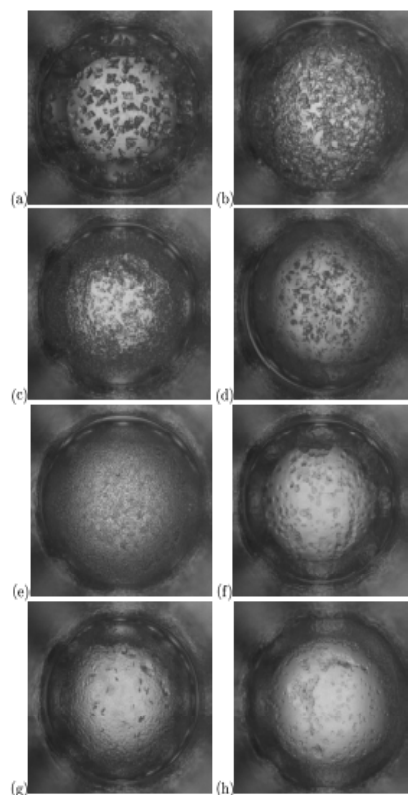


FIGURE 5.23: There are many diverse forms and shapes of microcrystals, which makes their identification challenging.

Figure 1 Diverse crystal forms.

Image classification challenge

Individual images have to be first analyzed to determine their morphologic features, and then use combination of these features to classify them into a predefined set of categories, as shown in Figure 2.

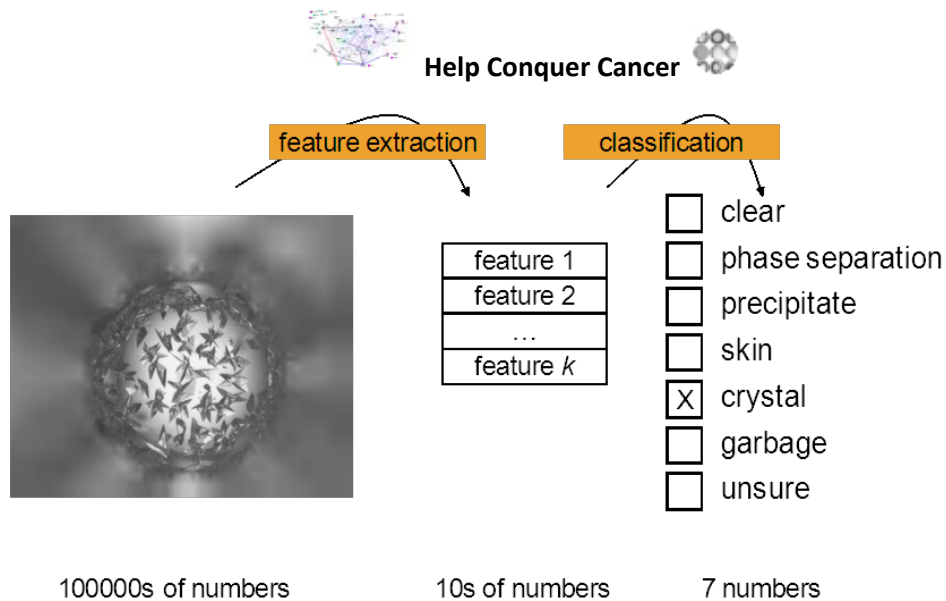


Figure 2 Image classification process.

Phase 1

During the first phase of the project, we processed a well-characterized set of images. We are in the process of optimizing features across wide range of individual image categories.

- **Truth data set:** 165,416 images, classified by 3 experts into 10 categories;
- **Image analysis:** 12,375 features on 90 million images.

Using the WCG computing capacity, all feature extraction for hand-scored image data has been completed by January 2008. While we are determining the best subset of features to use, we continue feature extraction for all unscored data. Our preliminary results show that this comprehensive approach is useful and necessary since the relationship among features, their parameters, and image classes is not linear. Although it is not practical to wait till 2013 to compute features for all 86 million images, the only sensible option is to determine the best feature subset on a well-characterized set of hand-scored images, which covers a large number of diverse image classes. This computation has been finished in January 2008, enabling us to perform the following feature selection:

- Assess the value of all computed features;
- Compute information content of single features (mutual information between feature & image class);
- Calculate information content of feature pairs (mutual information between feature-feature pair & image class);
- Compute information content per CPU-second (feature utility);
- Evaluate information density across feature-parameter space;
- Select “optimal” subset of features for image classification, i.e., the most informative features, with preference for less computationally intensive features when possible without decreasing accuracy.



Computed features will be used to identify essential combination of features that will lead to accurate image classification (such as a 3-class classification in Figure 3), i.e., classification that achieves both high sensitivity and specificity.

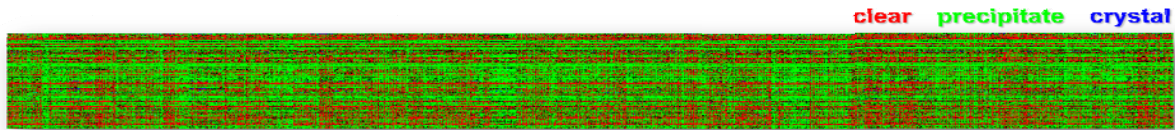
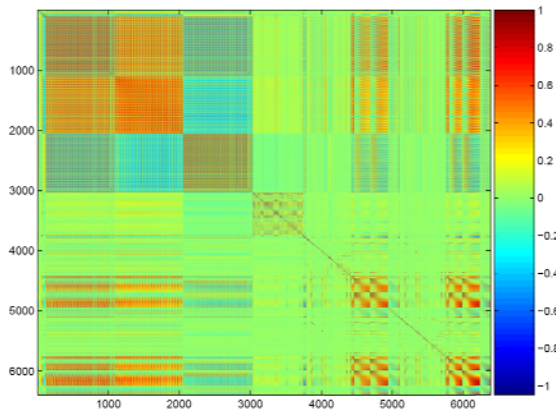


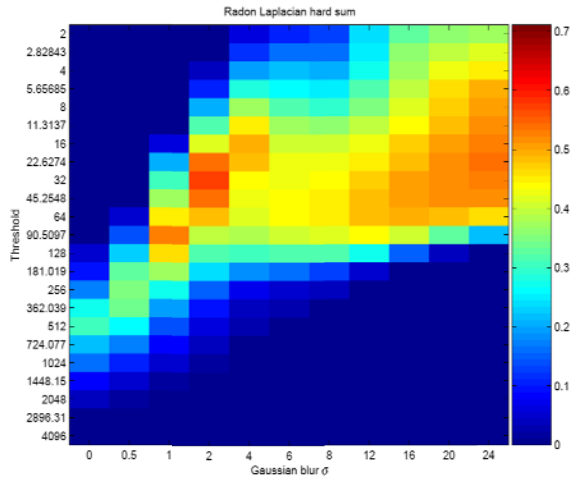
Figure 3 Image classes across the truth data set.

As shown in the examples below, we can optimize which features and which parameters are useful for predicting image class (see Figure 4, 5). But the process is challenging as features and parameters highly depend on the image class. Thus, we need to consider feature and parameter optimization per class as well, as shown in Figure 6.



Optimizing the feature set

Figure 4 Correlation of individual features across all image categories



Optimizing parameters

Figure 5 Optimizing parameters for individual features across all image categories. Heat maps indicate mutual information (measured in bits) between features (plotted in parameter space).

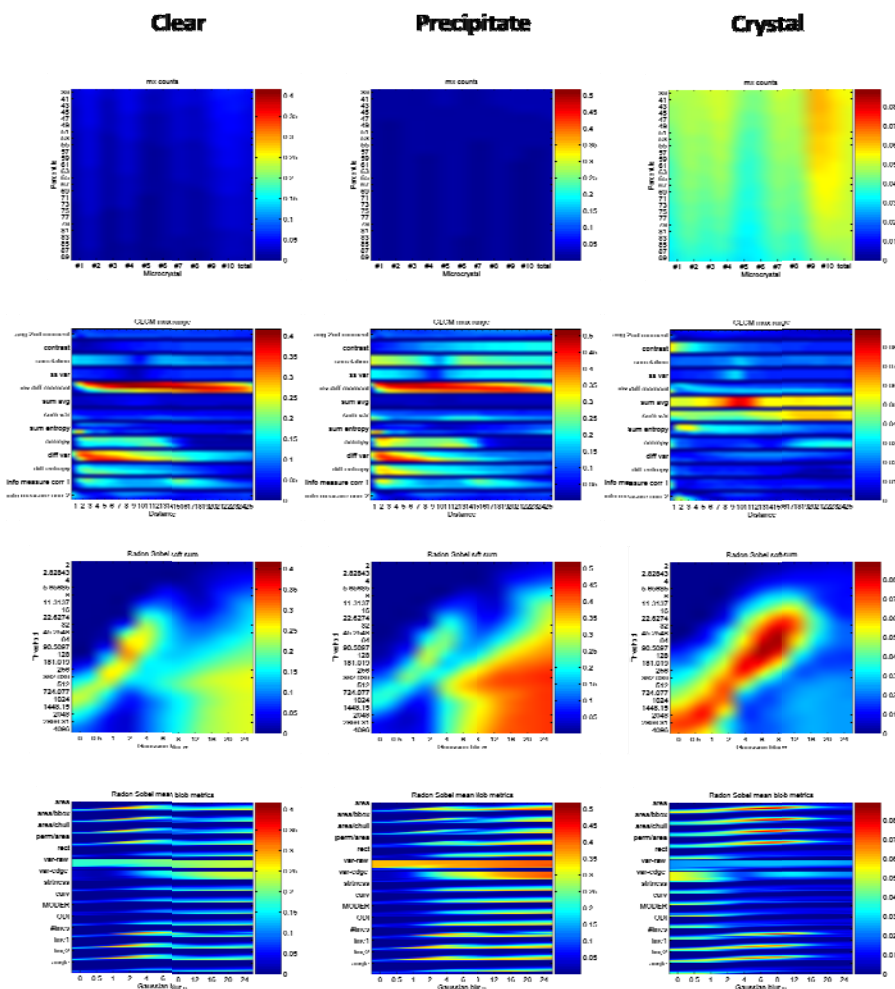


Figure 6 Effect of parameter changes to the information contents of image features. Heat maps indicating mutual information (measured in bits) between features (plotted in parameter space) and specific crystallization outcomes (clear, precipitate, crystal) is shown. Note how different regions of each feature family’s parameter space are sensitive to different crystallization outcomes. Peaks in these plots indicate candidate features for HCC Phase II.ⁱⁱ

Preliminary image classifiers

We have used a set of handpicked 74 features from peaks in the clear, precipitate and other mutual information plots to built two preliminary classifiers, using a Naïve Bayes model:

- **three-way:** clear, non-crystal precipitate, other;
- **ten-way:** clear, phase separation, phase + precipitate, skin, phase + crystal, precipitate, precipitate + skin, precipitate + crystal, crystal, garbage.

Using the training set of images and a leave-one-out cross-validation, we have measured how accurate each classifier is in identifying image from individual categories, i.e., what is the sensitivity and specificity of each classifier.

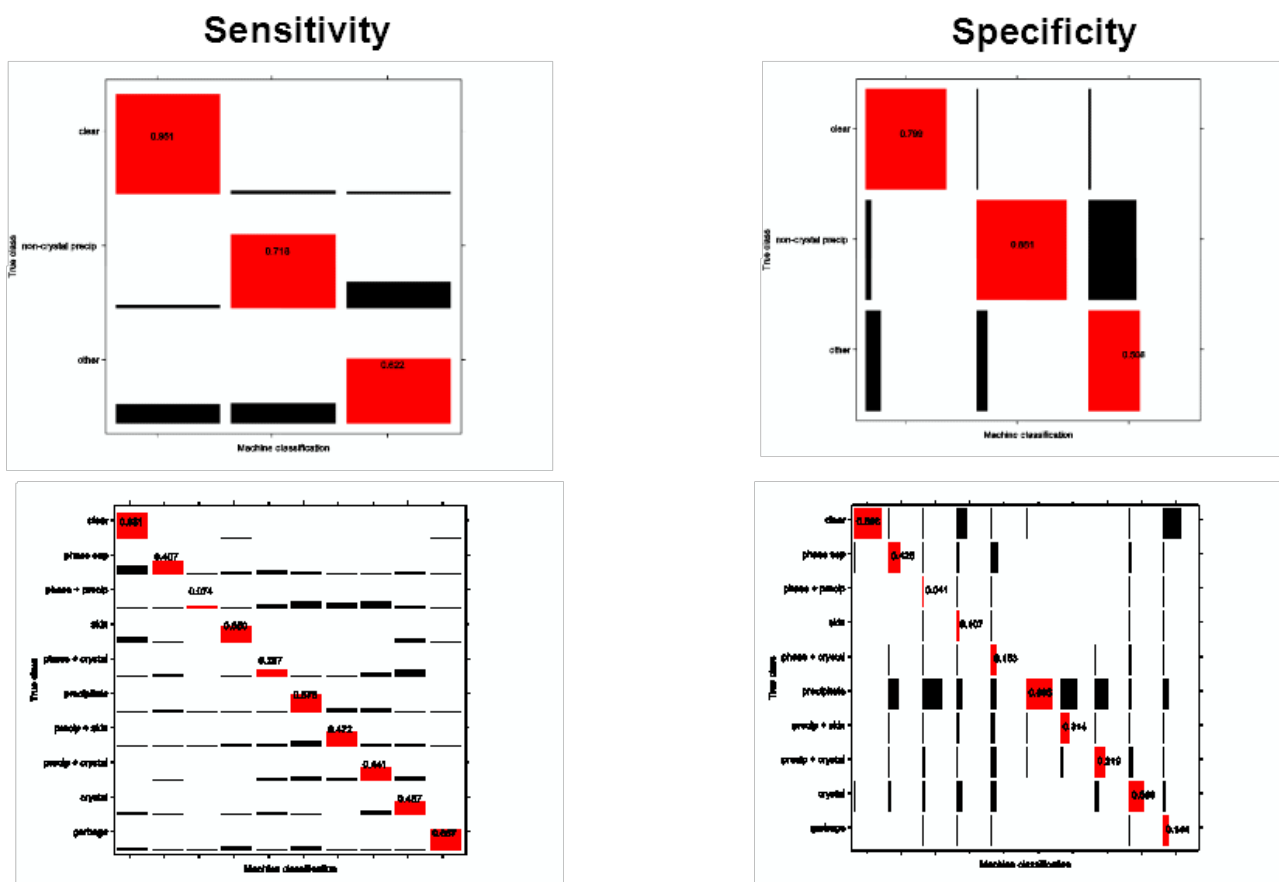


Figure 7 Naïve Bayes classifiers for 3 and 10 classes.

Future directions

- Improve image analysis to achieve high specificity and sensitivity in multi-class experiment categorization, and improve scalability to near real time.
- Protein crystallization principles derived from the crystallization database by data mining.
- Identify potentially successful conditions for proteins that were not yet crystallized.
- Crystallization optimization plans derived by combining case-based reasoning system and data mining.

As a result, more structures will be determined for larger number of important cancer proteins.

Thank you,

C. A. Cumbaa and I. Jurisica

¹Jurisica, I., D. A. Wigle. *Knowledge Discovery in Proteomics*, Mathematical & Computational Biology Series, Volume 8, Chapman & Hall/CRC Press, 2006.



ⁱⁱ Cumbaa, C. A., and I. Jurisica. Crystallization image analysis on the World Community Grid. *NIH PSI Bottlenecks Meeting*, Bethesda, MD, March 2008.