Update October 2008

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

We do benefit from this computing power greatly, but we also participate in WCG as an Integrative Discovery Team:  .

Since the launch of Help Conquer Cancer project in November 2007, WCG members contributed almost 19,627 years of run time, averaging about 50 years a day. To date 24,502,692 results were returned, completing about 16% of the work (*Statistics Last Updated: 11/20/08 12:06:03*).

Background

A crystallography experiment begins with a well-formed crystal that diffracts X-rays to high resolution. Despite the progress in experimental techniques, it remains challenging to crystallize novel proteins, and we still do not know most of the laws by which proteins adopt their three-dimensional structure. Thus, understanding these laws is one of the primary challenges in modern molecular biology.

While many steps in the pipeline from protein production to structure determination can be optimized, the crystal growth remains the rate-limiting step: as of August 21, 2008, the NIH Protein Structure Initiative (PSI) structural genomics centers reported 22,875 purified and soluble macromolecular targets, but only 7,804 (34%) of them were crystallized. Further, Protein Data Bank includes structures for only 50,000 proteins, and thus many disease-related proteins remain uncharacterized.

Results

Training/Testing Data Set

As a first step toward automated image classification colleagues at HWI have manually classified 147,456 images representing crystallization experiments from 96 different macromolecular samples[1]. Each image has been classified by three experts into seven predefined categories, or their combinations. The resulting data, where all three observers are in agreement, provides one component of a truth set for the development

[1] Snell EH, Luft JR, Potter SA, Lauricella AM, Gulde SM, Malkowski, M.G., Koszelak-Rosenblum M, Said MI, Smith JL, Veatch CK, Collins RJ, Franks G, Thayer M, Cumbaa C, Jurisica I, DeTitta GT. Establishing a training set through the visual analysis of crystallization trials part I: ~150,000 images. *Acta Crystallographica* ***D*** 2008. In press.

and rigorous testing of automated image classification systems, and provides information about the chemical cocktails used for crystallization.

To complement a training set from the visual observation of 147,456 crystallization outcomes a set of crystal images was produced from 106 and 163 macromolecules under study for the North East Structural Genomics Consortium (NESG) and Structural Genomics of Pathogenic Protozoa (SGPP) groups, respectively[2]. These crystal images have been combined with the initial training set.

Image Classification

From the crystallography viewpoint, it is not as important to find all crystallization results for a given protein (i.e., to identify all crystals from the set of 9,216 images); rather, the goal is to identify conditions that may lead to successful crystallization. Our preliminary analysis shows that a crystal hit (if present) is found in the first three hundred top-scoring images on a plate. 95% have a crystal in the top 100 images and 74% of all plates have a crystal in the top 10 identified images.

As an initial study, we constructed a classifier for separating images into 3 and 10 classes: crystal, crystal + phase separation, crystal + precipitate, precipitate, precipitate + skin, precipitate + phase separation, phase separation + skin, phase separation, clear, garbage. The classifier was trained and tested on a set of 88,562 unanimously scored images belonging to those classes, in a 10-fold cross-validation study. Each image was pre-analyzed by the “generation 4” image analysis system (840 features, a core subset of our current feature set).

The classifier is a hierarchical ensemble of 40 Gaussian mixture models (GMM), four models per class. Each model approximates the distributions of class-positive and class-negative images (e.g., those images depicting crystal + precipitate vs. those images depicting any other outcome) in a multi-dimensional feature space as two clouds of points with a multidimensional Gaussian probability distribution. The image features of each GMM (subsets of size 8, 16, 24, or 32 of the original 840 computed in the image analysis process) were chosen by a simulated-annealing algorithm as the subset of features that best discriminate the (e.g., crystal + precipitate) class from non-members of the class.

A new image is classified by feeding its features to the 40 GMMs, computing from each the likelihood of the image belonging to that GMM’s class’s distribution. Likelihoods for the 4 GMMs in each class are summed, and the sums for each of the 10 classes give the relative probability of the new image being a member of any one class.

---

[2] Snell EH, Lauricella AM, Potter SA, Luft JR, Gulde SM, Collins RJ, Franks G, Malkowski MG, Cumbaa C, Jurisica I, DeTitta GT. Establishing a training set through the visual analysis of crystallization trials part II: Crystal examples. *Acta Crystallographica* ***D*** 2008. In Press.

The results show improved detection of crystals as a whole (69% sensitivity, 78% specificity when combining crystal, crystal/phase, crystal/precip classes); individually, much of the confusion of those 3 classes lies between each of them, with the rest falling between crystals and phase, crystals and clear. Confirming earlier experiments, precipitate and clear drops are well-modeled (achieving specificity of 86% and 89% respectively with a preliminary Bayesian classifier). Phase separation alone is well detected, but precipitate + phase separation, phase separation + skin are poorly modeled, mostly due to small number of examples available in the initial training.

The 10-fold cross validation on the set of 88,562 images will be used to build the weights for the ensemble classifier, and to validate the results. The remaining 76,854 human expert classified images will then be used as external validation of the classifier. We will take the consensus among three experts plus a computer classification, and resolve discrepancies by further manual evaluation by experts.

Validated classifier will be applied to all 94 million images, to create a database of crystallization results, and to identify previously missed successful crystallization screens. Thus, we anticipate to significantly increasing the current rate of 34% successful crystallizations from PSI screens.

In the preliminary analysis (with neither feature space, nor the ensemble classifier optimized) of 1,580,544 images never classified by human expert coming from non-crystallized proteins, we have been able to determine ~53% images with crystallization leads, and ~7% with high-confidence hits (108,736). Manual inspection of a subset (~500) of these results determined: 27 microcrystals, 22 medium/large crystals, and 26 possible crystals. False positives included: 119 microcrystal precipitates or other very-fine precipitates, 73 heavy skin or skin/fine precipitates, and 241 medium/coarse precipitates and/or phase separation.

Future directions

We anticipate the following results: **(1)** rationally optimized set of image features and significantly improved crystallography image classification, as measured by improved both sensitivity and specificity; **(2)** re-analysis of 95 million images from 11,000 proteins is likely to discover proteins with favorable crystallization conditions that were missed in the initial human inspection (as only one of the six time-points per protein was ever assessed manually); **(3)** data mining of the resulting unified database of 15 million successful and failed crystallization trials will provide useful inside into chemistry about proteins, across time and crystallization conditions; **(4)** this unique crystallization database will provide an invaluable resource for optimizing crystallization plans for novel proteins. Making it publicly available will also enable other scientists to mine this rich resource.

Considering that HWI screens over 200 proteins a month in average from a wide scientific base, combined results from this research will provide a unique high-throughput pipeline from fast search phase, through automated image classification, to

lead discovery either by identifying crystals or by using data mining and case-based reasoning. Our results will directly positively impact PSI pipeline, and by making the classifier and data mining results publicly available, we will reach not only crystallographers, but biologist at large. Increasing crystallization success rate will thus lead to structure determination of larger number of important proteins.

Thank you,

C. A. Cumbaa and I. Jurisica