Automated Scoring of Crystallization Trials

Roy Liu[1], Yoav Freund[1], and Glen Spraggon[2]

[1]Computer Science Department, University of California at San Diego
[2]Genomics Institute of the Novartis Research Foundation

Recently, the use of robotics and parallel techniques for protein production and crystallization is becoming commonplace among structural genomics initiatives, whose contributions amount to 73% of newly solved structures each year. Despite the strides made in increasing physical experimental throughput, the act of finding just a few crystals among potentially thousands of crystallization experiments still remains a task for humans. A number of processes have been proposed for this task [6, 2, 5, 4, 7] and achieve varying degrees of success. Whilst automating the analysis part of a pipeline may seem like a straightforward task of recognizing lines and textures indicative of crystals, devising an automated analyzer in practice proves challenging for two reasons. First, computer vision is still a relatively young field. While many consider detection of ubiquitous, structured objects like human faces a well-studied problem, detection of non-uniform objects like crystals remains open and problem specific. Second, the needle-in-a-haystack property of finding just a few harvestable crystals among potentially thousands of trials necessitates a low false negative rate traded off against a tolerable false positive rate.

Our work attempts to address the above two challenges of the automation problem through a scoring based system – machine learning algorithms assign a score, or real-valued likelihood, of containing crystalline material to each trial. Specialists then look through the trials in rank-order to determine candidates for diffraction analysis. The proposed scheme bears a passing resemblance to previous works [2]; however, the authors there do not explicitly describe a ranking centric system. Consequently, we focus exclusively on a scoring framework and distinguish data sets as separate experimental attempts to crystallize distinct proteins.

The trained algorithm scores square image subregions of 127×127 pixels; the score for an entire image is the maximum over all square scores, as in Figure 1. This is not unlike previous work [5, 4] that also eschews global heuristics in favor of accurate local classifiers. Feature extraction relies on Gabor wavelet responses to detect edges and textures [5]. Orientation histogram statistics are also calculated and substitute for gray level co-occurrence matrices [6, 4].

To learn from extracted features, we use the alternating decision tree variant of boosting [3, 1]. Taken as a black box learning algorithm, boosting has the same input-output interfaces as support vector machine [5] (SVM), linear discriminant analysis [4] (LDA), and neural networks [6]. We choose boosting over other techniques for its ability to automatically combine many marginally discriminative features into a single, highly accurate ensemble classifier. Our choice seems timely in lieu of recent work on ensemble classification [4, 7] where the authors merge the outputs of two techniques into a single classification. Consequently, we view boosting as a principled, theoretically justified next step along these lines.

We report the scoring results of 319,112 crystallization trial images constituting the data sets of all 150 structures solved by the Joint Center for Structural Genomics during the 2006-2007 year. Our system achieves a mean ROC-AUC score of 0.918 averaged over set scores. Simulations indicate an expected 94% savings in human effort when searching, in rank-order, for the first instance of each set that yielded an x-ray crystal structure. Alternatively, a hypothetical cutoff accepting the top 25% ranked trial images of each set and rejecting the rest would have captured at least one structure-yielding instance for 143 out of 150 sets. These results suggest that computer assisted analysis can augment, rather than require modifications to, existing image-based crystallization systems; ultimately, they may provide full annotation of crystallization, thus enhancing our ability to record crystallization results and derive optimal crystallization conditions for specific proteins.

Please see http://hubris.ucsd.edu/paper.pdf for a paper in progress detailing the measurements and methods used.

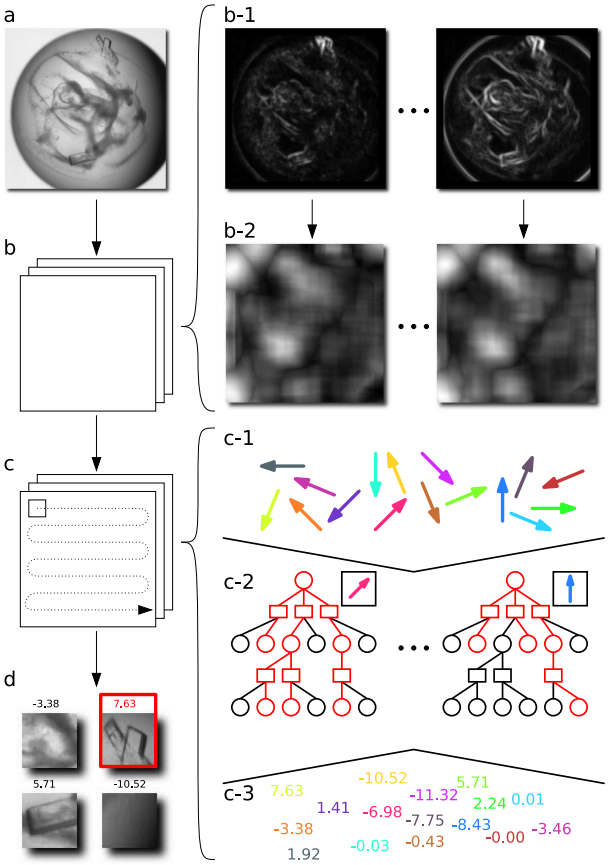

Figure 1: The scoring pipeline. **a)** the original image; **b)** an image stack obtained from image processing; **b-1)** heatmaps of Gabor responses. White areas represent pixels of greatest response; **b-2)** heatmaps of orientation histograms. White areas represent represent square centers with greatest "largest bin" statistic; **c)** scanning the image and scoring each square; **c-1)** each square is associated with a feature vector; **c-2)** each feature vector propagates differently through the alternating decision tree. The nodes marked in red contribute their weight to the total score; **c-3)** a real-valued score is associated with each feature vector; **d)** the maximum score over all squares doubles as the image score. The square with maximum score is marked in red.

References

[1] Jboost. http://www.cs.ucsd.edu/~aarvey/jboost/.

[2] Igor Jurisica Christian Cumbaa. Automatic classification and pattern discovery in high-throughput protein crystallization trials. *Journal of Structural and Functional Genomics*, 6:195–202, 2005.

[3] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 124–133, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[4] Kuniaki Kawabata, Kanako Saitoh, Mutsunori Takahashi, Mitsuaki Sugahara, Hajime Asama, Taketoshi Mishima, and Masashi Miyano. Integrated state evaluation for the images of crystallization droplets utilizing linear and nonlinear classifiers. *Acta Crystallographica Section D*, 62(9):1066–1072, Sep 2006.

[5] Shen Pan, Gidon Shavit, Marta Penas-Centeno, Dong-Hui Xu, Linda Shapiro, Richard Ladner, Eve Riskin, Wim Hol, and Deirdre Meldrum. Automated classification of protein crystallization images using support vector machines with scale-invariant texture and gabor features. *Acta Crystallographica Section D*, 62(3):271–279, Mar 2006.

[6] Glen Spraggon, Scott A. Lesley, Andreas Kreusch, and John P. Priestle. Computational analysis of crystallization trials. *Acta Crystallographica Section D*, 58(11):1915–1923, Nov 2002.

[7] Christopher G. Walker, James Foadi, and Julie Wilson. Classification of protein crystallization images using fourier descriptors. *Journal of Applied Crystallography*, 40(3):418–426, Jun 2007.