Data Mining and Machine Learning for Improved Crystallization Success - Expectations and Reality

Bernhard Rupp, *University of California - LLNL, L-448, POB 808, Livermore, CA 94551, USA.* E-mail: br@llnl.gov

Protein crystallization has traditionally been viewed as an art, with the outcome largely dependent on the skills (or superstitions) of the experimenter and a good portion of luck. Many competing 'recipes' for improvement include largely anecdotal and singular evidence, praying on the desperation of the unlucky experimenter.

The advance of automated protein crystallization methods over the past several years now provides the opportunity to amass substantial amounts of crystallization data. Direct capture of all experimental conditions and outcomes - including negatives - into relational data bases should in principle allow data mining and machine learning in the hope to unearth statistically valid knowledge about how to select and optimize the best crystallization conditions for a given protein.

Indications have emerged that the process of knowledge-based predictions in protein crystallization is not going as smoothly as one would hope. The foremost reason lies in the complex and locally determined nature of the crystallization process, and the high dimensionality and sparse sampling of the multivariate crystallization parameter space [1]. Optimal experimental design, careful annotation, and robust machine learning methods have provided various reliable general rules - often affirming prior empirical suggestions - while specific predictions yet remain of limited statistical significance due to low confidence of the derived rules.

[1] Rupp B., Wang J., *Methods*, 2004, **34**, 390-407

Keywords: crystallization, data mining, predictive models