ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Structural Biology 142 (2003) 162–169

Journal of Structural Biology

www.elsevier.com/locate/yjsbi

Maximum-likelihood crystallization

Bernhard Rupp*

Macromolecular Crystallography and TB Structural Genomics Consortium, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

Received 5 February 2003

Abstract

The crystallization facility of the TB Structural Genomics Consortium, one of nine NIH-sponsored structural genomics pilot projects, employs a combinatorial random sampling technique in high-throughput crystallization screening. Although data are still sparse and a comprehensive analysis cannot be performed at this stage, preliminary results appear to validate the random-screening concept. A discussion of statistical crystallization data analysis aims to draw attention to the need for comprehensive and valid sampling protocols. In view of limited overlap in techniques and sampling parameters between the publicly funded high-throughput crystallography initiatives, exchange of information should be encouraged, aiming to effectively integrate data mining efforts into a comprehensive predictive framework for protein crystallization.

Keywords: High-throughput crystallization; Statistical analysis; Crystallization screening; Structural genomics; Predictive models

1. Introduction

The National Institutes of Health (NIH) funding for a substantial Structural Genomics (SG) Initiative (Norvell and Zapp-Machalek, 2000), under way since fall of 2000, provides significant public funding to nine P50 Structural Genomics Centers. One of the main objectives of these centers is the advancement of high-throughput crystallography, including methods for high-throughput protein crystallization. Two major conclusions can be drawn already at this early stage from initial communications from the public efforts. First, it has become obvious that a bottleneck potentially more significant to high throughput than protein crystallization is the production of proteins (in particular, ''inherently crystallizable'' proteins; Segelke, 2001) and second, that a number of different crystallization methods can reasonably be adapted to achieve high throughput. Few reports are yet available on new crystallization statistics and predictions from the initiatives, and a distinct probability exists that under pressure to produce structures (which is the ultimate goal of a SG Center), the opportunity to create comprehensive and consistent crystallization databases through the centers may be lost. This concern is not entirely unfounded, as both the omission of negative results and the lack of the most basic quantity in statistics, the number of trials, have made the publicly available databases (Biological Macromolecule Crystallization Database or BMCD, Gilliland et al., 1994; Protein Data Bank or PDB, Berman et al., 2000) very difficult to use for the purpose of rigorous statistical analysis and inference without significant restructuring and annotation (Hennessy et al., 2000). This brief report discusses general aspects of experimental design of crystallization experiments in view of statistical analysis and machine learning, interspersed with implementation and preliminary results obtained at the crystallization facility of the TB Structural Genomics Consortium (TBSGC; Goulding et al., 2002), one of the nine NIH-funded structural genomics pilot projects. A major objective of the TBSGC crystallization facility has been to maximize overall operational efficiency within the budget constraints of public funding. An overview of our underlying philosophy and the challenges faced in the first 2 years of the TBSG crystallization facility is provided in a separate review (Rupp, 2003). Technical implementation details, including crystallization robotics, crystal recognition, data collection, and structure solution, are provided in the special literature (Rupp et al., 2002).

* Fax: +925-424-3130.
E-mail address: br@llnl.gov.

1047-8477/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1047-8477(03)00047-9

2. High-throughput crystallization

In a high-throughout environment, large numbers of samples, with limited prior knowledge of each, need to be processed. A compromise is necessary between effort, time, and material spent on in-depth physicochemical analysis of the crystallization process of each specific protein leading to rational approaches directed toward improving crystallization (Dale et al., 1999; D'Arcy, 1994), while still achieving high throughput and the required high *overall success rate* for the effort. This is a most distinct difference from academic, strongly hypothesis-driven research, in which the determination of a single recalcitrant structure may be of utmost importance, in particular for the unfortunate graduate student engaged in the project. Many of our efficiency considerations do, however, apply to operations on any scale, and we believe that the procedures and instrumentation we have developed and use at the TBSGC crystallization facility are valuable for academic laboratories as well (Rupp, 2003).

2.1. Protein crystallization as a sampling problem

Let us conceptualize the "crystallization space" as an n-dimensional vortex (or hypercube of dimensionality n), whose bases (axes) are extensive parameters like chemical components and protein concentration, and intensive ones like temperature, setup technique, or pH. Crystallization success analysis can then be treated as a sampling problem of an unknown distribution of success rates in crystallization (parameter) space. Within one given system, the sampling can be done consistently and be statistically valid, and a variety of machine learning methods can be used to make inferences about the frequency distribution of successes. Comparing or merging results from a different sampling space (consortium or group), however, may be difficult or even impossible. For example, the crystallization technique employed may be a more critical parameter (in terms of factorial analysis) than the reagent basis set; and of the reagents used, the overlap between the sampling spaces may be poor. Fig. 1 uses Venn diagrams to visualize this problem. In the extreme case we are faced with a dilemma equivalent to merging and scaling two diffraction data sets with no or poor overlap. The varying methods and approaches used by different consortia and groups will make it quite difficult to accurately compare success rates. Each of the approaches appears to work—which is to a degree expected from general success rate analysis, as we will see shortly—but consistent data mining will be difficult.

2.2. Sampling strategies in crystallization space

Given the unlimited number of combinations of components in crystallization recipes, it comes as no surprise that, historically, crystallization conditions were often chosen on the basis of what had worked before and what was available on the reagent shelf. Screening kits based on previous success analysis have been quite successful (Jancarik and Kim, 1991), and numerous variations of this first kit are now commercially available. In a statistical sense, repeated use of such premixed "sparse matrix" solutions amounts to oversampling of certain spots in the multidimensional crystallization

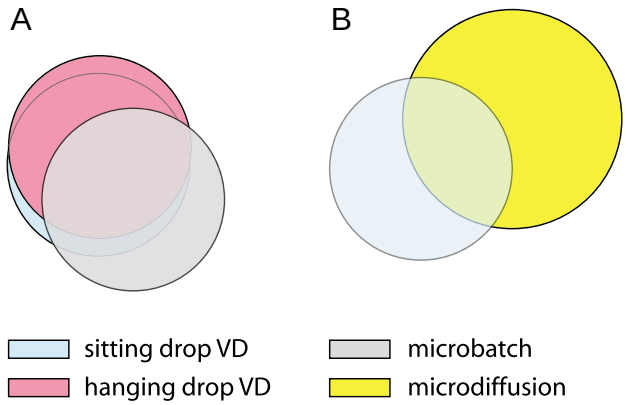

Fig. 1. Venn diagrams representing crystallization scenarios. (A) Overlap in success space between three different techniques. Circles have the same diameter, indicating equal overall success rate for each method. Overlap between hanging- and sitting-drop vapor diffusion (VD) techniques is presumably large, whereas microbatch may have fewer conditions in common with either. (B) Hypothetical scenario representing free interface diffusion microtechnique with higher success rate but limited overlap with sitting-drop VD technique. Similar diagrams can be used to visualize the overlap (or lack thereof) of basis sets used in different setups.

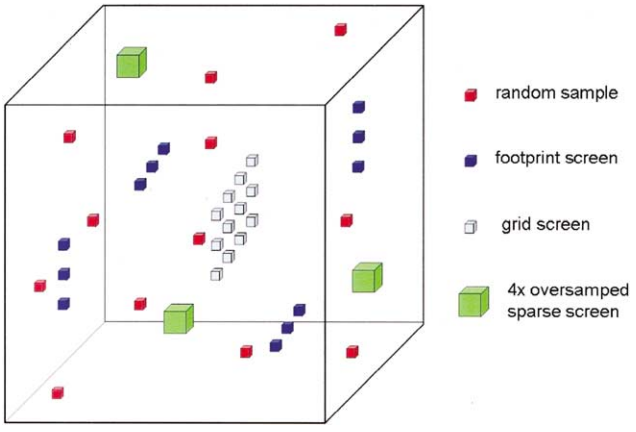

Fig. 2. Schematic of simple 3-D crystallization space showing varying coverage by different sampling protocols using 12 trials each. The large cube represents fourfold oversampling in a sparse matrix-type experiment. Note that grid screening is not only used to comprehensively screen for conditions, but also deployed with a rationale to explore systematic variations of two dimensions considered major factors while keeping other parameters constant (McPherson, 1982). Grid screening thus can be viewed as a complete pseudo-factorial on a 2-D subset of the sample space.

space. Reported success rates thus are limited to few specific combinations of a preselected basis set of reagents. Although the success of these screens is demonstrated (and expected, as we will argue shortly), there are serious shortcomings in this limited approach, affecting comprehensiveness and the predictive power of statistical analysis.

Probabilities and intuition. While humans excel in complex visual tasks like pattern (crystal) recognition, our intuition regarding randomness and conditional probabilities[1] is generally not as well developed. Assume a sequence of five heads in a row upon the throw of a fair coin showing either head (H) or tail (T). Although perplexing, we can understand that the probability for event HHHHH is not absurdly unlikely, and, that in the *long run*, it may well turn out that the coin is fair, i.e., $p(\mathrm{H}) = p(\mathrm{T}) = 0.50$. We may also reason under consideration of a strong prior, namely that a coin showing only heads is an unlikely proposition. In the case of limited numbers of crystallization trials and no prior knowledge about success probabilities, there is very little to guide us in terms of what presents a true "hot spot" in crystallization space and what will become insignificant in the long run.[2] Sequential analysis and corresponding redesign of experiments based on premature analysis of a limited sample set leads to self-fulfilling prophecies misguiding the search for potential true hot spots in crystallization space. The problem is compounded by the fact that a significant number of proteins crystallize quite easily and under multiple different conditions (Segelke, 2001). It is therefore even *expected* that a not unsubstantial success rate can be achieved regardless of which screening method is used and what reagent basis set is used. A brief discussion of a Bayesian framework applied to crystallization data analysis will illustrate this point.

Statistical approaches to sampling. In view of the above-mentioned limitations, the need for a rigorous statistical approach to crystallization screening and optimization was recognized early by Carter and Carter (1979), who suggested factorial designs, the use of variance analysis, and response surface optimization (Carter, 1999). Factorial designs attempt to balance the occurrence of factors and of their pair-wise combinations during the sampling process. Optimal implementation of such methods suitable for comprehensive analysis requires specific cocktails to be prepared for screening and optimization, and unfortunately, when these advanced concepts were first introduced, availability of robotics was not as widespread as it is becoming now (a major factor contributing to the widespread popularity of prefabricated kits).

In a seminal paper, Segelke (2001) has assessed various crystallization screening protocols (Jancarik and Kim, 1991; McPherson, 1982; Stura et al., 1992) in terms of sampling efficiency, i.e., finding crystallization conditions with a minimum number of trials (Fig. 2). Based on formal statistical derivation, Segelke has shown that random (combinatorial) sampling is most efficient, particularly when *success rates are low or clustered*. His efficiency analysis also allows estimating the number of trials above which return on investment (time, supplies, and protein) during further screening diminishes, as indicated by cumulative probabilities (Segelke, 2001, Fig. 4). For the average soluble protein, based on available frequency and success rate data, we estimate that 288 (3×96) trials should suffice to find crystallization conditions with high probability. Beyond this point, the options of protein engineering (examples are discussed in Dale et al., 1999; Edwards et al., 2000; Waldo et al., 1999) and search for orthologs should be investigated, aiming to obtain an inherently more crystallizable variant (Segelke, 2001) of the particular protein.

2.3. Bayesian considerations

In random sampling, coverage of the crystallization space is achieved by using each crystallization condition only once. At the same time, in contrast to factorial designs (Carter, 1999), no assumptions about any success rate distributions or about factors specific for a particular protein are made.[3]

The omission of prior knowledge or absence of assumptions may seem a serious limitation of the random screening, but there is good reason not yet to deviate from the assumption of ignorance. As already indicated in the discussion of sparse matrix sampling, the inclusion of prior knowledge—either consciously during analysis or, quite insidiously, inherent in the experimental design—may affect the outcome of the estimate of the posterior, in our case crystallization success probabilities. The formal basis for this reasoning is given by Bayes' theorem, which can be derived from the sum and product rules of conditional probabilities (for an introduction see Sivia, 1996). In simplified form and assuming a constant evidence term,

[1] An excellent review describing classical examples of misperception of randomness and conditional probabilities is contained in Bennett (1998). The examples include the well-known jailer's paradox and the varying gender probability for offspring given different prior knowledge of the birth order and the gender of siblings. For an example demonstrating consequences of focusing on single events (similar to crystallization successes) without maintaining negative controls, see http://sprott.physics.wisc.edu/pickover/esp.html.

[2] I am particularly skeptical in this context of the proliferation of crystallization tips, which practically never undergo any serious statistical validation. One should be aware of "tips" particularly if they lack clear underlying rationale.

[3] Note, however, that the CRYSTOOL protocol (Segelke and Rupp, 1998) allows easy customization of parameter ranges such as pH, reagent concentrations, and reagent frequencies once clearly established trends emerge.

$$\text{prob}(C|D,I) \propto \text{prob}(D|C,I) \times \text{prob}(C|I). \quad (1)$$

Formula (1) essentially states that the posterior probability prob$(C|D,I)$ of crystallization success (C), given our data (D) and prior information (I), depends directly on the likelihood function prob$(D|C,I)$ *and* the prior probability prob$(C|I)$ of crystallization success without having analyzed the data. It is evident that any prior assumption, made directly or imposed indirectly through a suboptimal experimental design, will affect our success prediction. The most striking point is that the posterior recovers rather *slowly* from incorrect but plausible priors, whereas a uniform prior (or unreasonable prior) is more effectively overcome by the likelihood function (i.e., data).[4] The proportionality in Eq. (1) also shows that "bootstrapping" by using the resulting posterior probability as a prior (directly or through experimental design changes) and subsequent reanalysis leads to a much sharper posterior probability distribution function, i.e., a serious overestimate of our success rate (Sivia, 1996). Both findings are of great relevance to our crystallization experiments and should caution against early assumptions about success rate distribution in crystallization space as well as against a premature limitation of the basis set used in the experiments based on frequency analysis alone. An interesting project utilizing Bayesian analysis has been described by Hennessy et al. (2000), but no recent updates have been made available (http://www.xtal.pitt.edu/xtalgrow/default.htm).

2.4. Choice of parameter basis set

Most crystallization screen designs tend to preclassify reagents into one or more groups such as precipitant, additive, buffer, and detergent and to use a combination of one (or no) reagent from each of these classes. At present CRYSTOOL (Segelke and Rupp, 1998) uses 90 manually premixed stock solutions, divided into four groups: precipitants, buffers, additives, and detergents. About 50 different reagents form the chemical basis set, which can be expanded or updated if evidence suggests it. From simple preliminary frequency data (Fig. 3), it appears that the inclusion of detergents in the reagent basis set might be beneficial (as indicated by Cudney et al., 1994).

There has been a great deal of discussion about the relative virtues of chemicals comprising a basis set. Yet, whereas drop size, protein concentration, and other method-determined parameters may be as influential as the reagents used (Carter, 1999), systematic investigation of the nonchemical parameters spanning the crystallization space has received less attention. Drop size, for example, may significantly influence kinetics via nucleation rates (Bodenstaff et al., 2002). Regardless of what crystallization design is used, nonoverlap in crucial parameters, be they reagents or technical parameters, makes the comparison of results between different research groups rather difficult. This concern also holds for the protein classes processed, as results for small, highly soluble proteins will likely indicate overall success rates different from data including large complexes or membrane proteins.

2.5. Screening versus optimization

Frequently in the literature, a strong distinction is made between crystallization screening and optimization of crystal growth (for example, Chayen and Saridakis, 2002; Luft et al., 2001). From a chemical point of view, this distinction is rather arbitrary and mostly affects a change in method-related parameters, but those can significantly impact overall efficiency. Results from the TB facility indicate that about 1/3 of diffracting crystals grow directly out of random screens without any need for optimization. We expected this result based on the reasonably large number of trials (288) conducted for each protein and the assumption that a significant number of proteins crystallize easily and under multiple conditions and probably in multiple crystal forms (Segelke, 2001). The data so far indicate that about half of the proteins indeed form crystals in more than 5 of 288 different random conditions. We thus argued that it would be more cost effective to use a single technique (vapor diffusion in sitting drop IntelliPlates) for screening and optimization (Rupp, 2003), which enables us to use the same equipment throughout the process from screening to harvesting, and to automatically set up either grid-type or limited random-type optimizations designed with CRYSTOOL. Consistency in setup would also avoid possible nonoverlap when transitioning between different crystallization techniques.[5]

However, protein production appears to become the more serious challenge for achieving high throughput than crystallization setup itself, despite promising developments in protein engineering (Edwards et al., 2000; Waldo et al., 1999). Given 150–300 μl as the lower limit of protein required for a complete screening in our setup, a significant reduction in cost might be possible if microtechniques were used in combination with integrated micropurification. Excluding proteins which are unlikely to crystallize from expensive expression and

[4] A very convincing demonstration of this behavior of the posterior probability distribution function is given in Sivia (1996) based on a simulation of coin tosses.

[5] Analysis of 5000 experiments comparing five different sitting- and hanging-drop vapor diffusion setup techniques (Schick et al., 2001) indicated that although overall success rate does not significantly differ for the 11 proteins under 96 conditions tested, significant differences do exist for specific conditions under different techniques. We are not aware of any systematic studies conclusively demonstrating overall superiority of any one of the commonly used techniques (sitting drop, hanging drop, batch methods, etc.).

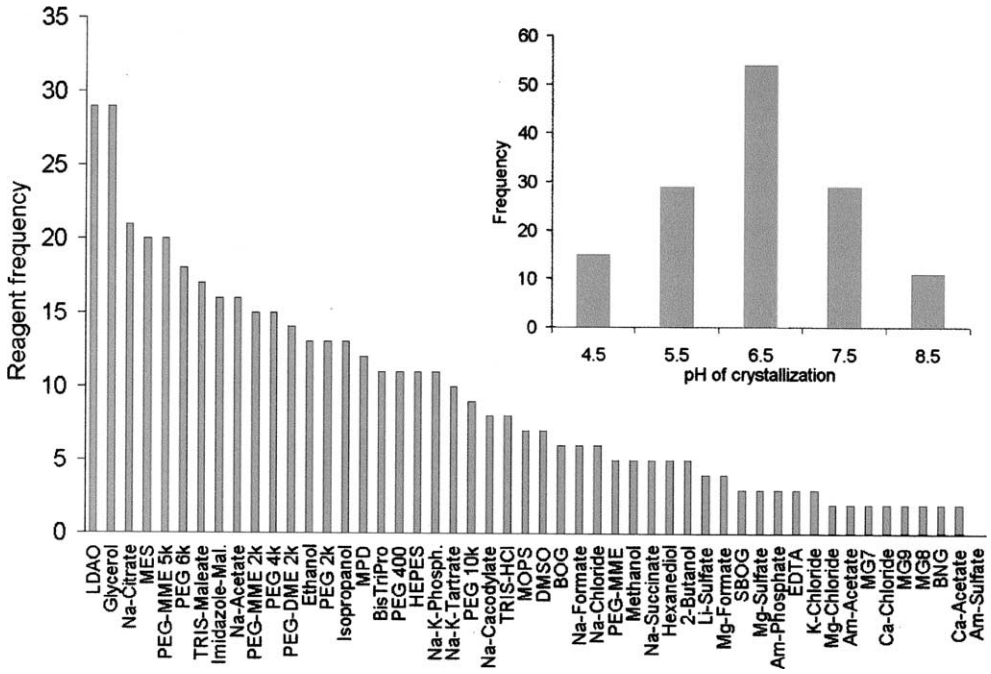

Fig. 3. Reagent frequency distribution in 203 successful random crystallization experiments and pH distribution of crystallization successes (inset). The detergent LDAO (lauryldimethylamine-N-oxide) and glycerol appear to be effective additives. The distribution of precipitants and salts appears within expectations, with the notable exception of no reported successes for ammonium sulfate, a popular precipitation agent (McPherson, 1999, p. 204), although it was included with equal probability in about 800 experiments. The pH distribution, although coarsely sampled, follows the distribution extracted from PDB data. Note that with only about 200 successful crystallization conditions forming a marginal sample one can expect that in the long term the frequency distributions will change. The lack of overlapping basis sets makes it unfortunately impossible to compare the performance of malonate (McPherson, 2001; Vilasenor et al., 2002) with our highest ranked precipitants. In addition, even with the common setup technique of sitting-drop vapor diffusion, varying factors of potential significance like drop size, incubation temperature, protein set, and experimenter skills would make it difficult to compare the data.

purification scale up would provide cost reduction in purification and thus provide an overall efficiency gain despite possible loss or omission of some proteins due to nontransferability to a different optimization and harvesting technique (Fig. 1B). Small droplet sizes, ease of miniaturization, and absence of additional sealing requirements favor methods like microbatch under oil for crystallization screening (Chayen, 1998; D'Arcy, 1994; Luft et al., 2001), while similar considerations make free interface diffusion microcrystallization based on multilayer soft lithography (see www.fluidigm.com; Hansen et al., 2002) potentially promising. Some commercial ventures (Stevens, 2000) employ large-scale, advanced nanodrop setups for both screening and growth, although drop size does significantly affect kinetic parameters, potentially limiting parameter overlap in success space. Merits and caveats of nanocrystallization are discussed in Bodenstaff et al. (2002).

3. Implementation details

3.1. Crystallization cocktail production

As a consequence of de novo cocktail design for each protein construct, a large number of crystallization cocktails need to be prepared for random screening and optimization. We thus implemented customizable random screen generation in the computer program CRYSTOOL (Segelke and Rupp, 1998) and interfaced it with a liquid-handling robot to automatically produce crystallization cocktails in 96-well format (Rupp et al., 2002). Production of de novo random screens is time consuming (20–40 min per 96-well cocktail block) and thus rate limiting in our high-throughput crystallization process. To balance the desired comprehensive coverage of the crystallization space with the throughput requirements, one could use each of the 288-condition random screen sets for more than one protein. Such modest oversampling does not compromise the validity of random sampling data, but it allows us to conveniently screen about 10–20 protein samples per day, substantially exceeding the current protein production rate.

3.2. Crystallization plate setup

Requirements for dispensing precision, volume, and speed differ substantially for cocktail production compared to the actual plate setup. While large-volume (milliliters) handling with modest speed and precision requirements suffices for cocktail setup, fast and very accurate (also in geometric terms) dispensing of small

(microliter to nanoliter) volumes is mandatory for crystallization plate setup. We thus decided, at the expense of full integration, to separate the plate setup from the cocktail mixing step. By augmenting a Hydra multichannel dispenser with a contactless, single-channel microsolenoid dispensing unit, we developed a relatively inexpensive robot that can rapidly set up the cocktail reservoirs and, without rearraying losses, dispense the protein (500 nl to 1 μl) into drop wells filled with precipitant aliquots (Krupka et al., 2002).

3.3. Crystallization plate considerations

For simplicity of handling and ease of harvesting we elected to use the same sitting-drop setup throughout for screening, optimization, and harvesting. We designed a suitable, SBS-compliant, 96-well plate for sitting drops, IntelliPlate, that specifically accommodates the needs of our high-throughput process. The plate has wide, elevated rims for reliable tape sealing and different well sizes to accommodate various drop sizes or additional cryobuffer during harvesting. Polished wells reduce sticking of crystals and support easy harvesting. Both well shape and optical properties are optimized toward automated image acquisition and recognition. A synopsis of our plate imaging and crystal recognition software CRYSFIND is provided in the TBSGC facility review (Rupp, 2003).

4. Analysis of experiments

At the lowest level of crystallization data analysis, one seeks to make inferences about hot spots in success rates by simple frequency statistics, resembling the way the sparse matrix set (Jancarik and Kim, 1991) was assembled. Frequency analysis can be carried out globally for all proteins or by categorizing proteins into groups with distinct properties. Important for the long run are correlations (generalizations) within the phase space and with known properties of the protein, i.e., priors that modify our choice of hypothesis (prediction) of where to find the highest likelihood for success given what we know already—the classical Bayesian strategy. Technically we are facing a classification problem in a poorly sampled space of high dimensionality, with all the associated difficulties (for a review of statistical learning methods see, e.g., Hastie et al., 2001). In addition, the validity of any statistical inference method greatly depends on suitable experimental design and the consistency of the databases. Both points are not trivial and only rarely explicitly addressed in the crystallization literature (Carter, 1999; Jurisica et al., 2001; Hennessy et al., 2000). For such reasons, results from undirected cluster analysis of the BMCD do not translate directly into useful crystallization strategies (Farr et al., 1998). Substantial annotation and restructuring are necessary to obtain meaningful subgroups and to incorporate the analysis into a predictive framework (Hennessy et al., 2000). Memory-based machine learning algorithms like case-based reasoning have shown promise (Jurisica et al., 2001), and we are currently exploring the performance of methods like market basket analysis (MBA), which essentially answer the question "what goes with what" and work well for even frequency distributions as in the random sampling case. One of the advantages of MBA is that the results are presented in the form of clear, readable association rules, whose meanings are obvious, and responses can be implemented immediately.[6] In the end, no matter how sophisticated the statistical analysis and data mining of crystallization space, any of the techniques will provide a only basis for increasing the probability of crystallization success (or increasing our degree of belief in it), but never guarantee success for any particular protein.

5. Outlook and suggestions for the future

From our initial operation, as limited as the analysis of results has to be at this time, we still can draw a number of conclusions and propose some considerations for future work. The concept of random sampling seems to perform at least as well other methods (about 1/3 of the TB proteins received could be crystallized robotically without any prior assumptions, tips, or tricks), and of those, about 1/3 diffracted beyond 2.5 Å without need for optimization. In addition, by consistent random sampling, we are accumulating comprehensive and valid samples of the crystallization space with respect to our reagent basis set and method. The preliminary reagent frequency distribution for successful conditions varies from expectations based on the (mostly kit-based) frequency data. Limitations of our preliminary analysis are the low number of trials, limited overlap with other basis sets, and use of only one technique (sitting-drop vapor diffusion). Moreover, simple frequency analysis does not discriminate uniqueness of successes, correlate multiple successes, or discriminate data clusters.

To optimally reduce to practice the inferences of advanced statistical analysis and machine learning will require a dynamic response through individual cocktail preparation. In random sampling we are already producing de novo cocktails for each sampling experiment, and the widespread availability of (relatively) affordable liquid handling robots will help establish this practice.

[6] Testimony to the usefulness of MBA was the elimination of an annotation error in the success rate data base used to extract Fig. 3. Na formate *and* DMSO formed a binary group in 12% of the cases with an 86% grouping probability (5.8 times more likely that random), caused by improper entry of an optimization experiment as a random screen.

We predict that the dominance of the rigid, kit-based sampling screens will fade as soon as comprehensive analysis methods favor screening or optimization strategies that require on-demand cocktail production.

We also believe that open discussion of crystallization procedures between the groups involved in public high-throughput data analysis could help to improve data consistency and to populate databases with sufficient overlap, allowing a large-scale data mining effort. Different groups use different methods and targets, and without a fair basis for comparison, it will be rather difficult to determine the absolute merits of the approaches. It also appears that integration of micropurification with methods of microcrystallization may provide an overall increase in efficiency in a high-throughput environment, despite the transferability issues during scale-up for optimization and harvesting.

Finally, we are not aware of any access capacity of crystallization robotics being used for systematic large-scale studies of factors such as temperature, setup technique, protein clustering, and batch variation. Small differences in success rate under different setup or environmental conditions require large sample sets to become statistically significant, and as a consequence, systematic studies of such effects are largely absent in the literature or limited to specific cases. Coordination between the efforts and the funding of additional projects that focus on systematic exploration of crystallization space, without the pressure of operating in production mode, could be a worthwhile investment.

Acknowledgments

I thank the members of the TB Structural Genomics Consortium crystallization facility team (B.W. Segelke, H.I. Krupka, T. Lekin, J. Schafer, and D. Toppani) for their contributions to the development of techniques described in this paper. K.A. Kantardjieff, CSUF, has provided assistance with statistical data analysis and manuscript revisions. The cloning and protein production facilities under J. Perry, C. Goulding, D. Eisenberg (UCLA), T. Terwilliger, M. Park, and G. Waldo (LANL) have supplied a steady flow of proteins. I thank Jim Sacchettini, Texas A&M University, for support during my sabbatical leave. LLNL is operated by the University of California for the U.S. DOE under Contract W-7405-ENG-48. This work was funded by an NIH P50 GM62410 (TB Structural Genomics Center) grant.

References

Bennett, D.J., 1998. Randomness. Harvard University Press, Cambridge, MA.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Bodenstaff, E.R., Hoedemaker, F.J., Kuil, E.M., de Vrind, H.P.M., Abrahams, J.P., 2002. The prospects of nanocrystallography. Acta Crystallogr. D 59, 1901–1906.

Ducruix Jr., A.., 1999. Experimental design, quantitative analysis, and the cartography of crystal growth. In: Giege, R. (Ed.), Crystallization of Nucleic Acids and Proteins, second ed.. Oxford University Press, New York.

Carter Jr., C.W., Carter, C.W., 1979. Protein crystallization using incomplete factorial experiments. J. Biol. Chem. 254, 12219–12226.

Chayen, N.E., 1998. Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques. Acta Crystallogr. D 54, 8–15.

Chayen, N.E., Saridakis, E., 2002. Protein crystallization for genomics: towards high-throughput optimization techniques. Acta Crystallogr. D 58, 921–927.

Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y., McPherson, A., 1994. Screening and optimization strategies for macromolecular crystal growth. Acta Crystallogr. D 50, 414–423.

Dale, G.E., Kostrewa, D., Gsell, B., Stieger, M., D'Arcy, A., 1999. Crystal engineering: deletion mutagenesis of the 24 kDa fragment of the DNA gyrase B subunit from *Staphylococcus aureus*. Acta Crystallogr. D 55, 1626–1629.

D'Arcy, A., 1994. Crystallizing proteins—a rational approach? Acta Crystallogr. D 50, 469–475.

Edwards, A.M., Arrowsmith, C.H., Christendat, D., Dharamsi, A., Friesen, J.D., Greenblatt, J.F., Vedadi, M., 2000. Protein production: feeding the crystallographers and NMR spectroscopists. Nat. Struct. Biol. Suppl. 7, 970–972.

Farr Jr., R.G., Perrymann, A.L., Samudzi, C.T., 1998. Re-clustering the database for crystallization of macromolecules. J. Cryst. Growth 183, 653–668.

Gilliland, G.L., Tung, M., Blakeslee, D.M., Ladner, J., 1994. The biological macromolecule crystallization database, version 3.0: new features, data, and the NASA archive for protein crystal growth data. Acta Crystallogr. D 50, 408–413.

Goulding, C.W., Apostol, M., Anderson, D.H., Gill, S.D., Smith, C.V., Yang, J.K., Waldo, J.S., Suh, S.W., Chauhan, R., Kale, A., Bachhawat, A., Mande, S.C., Johnston, J.M., Baker, E.N., Arcus, V.L., Leys, D., McLean, K.J., Munro, A.W., Berendzen, J., Park, M.S., Eisenberg, D., Sacchettini, J., Alber, T., Rupp, B., Jacobs Jr., W., Terwilliger, T.C., 2002. The TB structural genomics consortium: providing a structural foundation for drug discovery. Curr. Drug Targets Infect. Disord. 2, 121–141.

Hansen, C.L., Skordalakes, E., Berger, J.M., Quake, S.R., 2002. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. Proc. Natl. Acad. Sci. USA 99 (26), 16531–16536.

Hastie, T., Tinshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P.A., Rosenberg, J.M., 2000. Statistical methods for the objective design for screening procedures for macromolecular crystallization. Acta Crystallogr. D 56, 817–827.

Jancarik, J., Kim, S.-H., 1991. Sparse matrix sampling: a screening method for the crystallization of macromolecules. J. Appl. Crystallogr. 24, 409–411.

Jurisica, I., Rogers, P., Glasgow, J.I., Fortier, S., Luft, J.R., Wolfley, J.R., Bianca, M.A., Weeks, D.R., DeTitta, G.T., 2001. Intelligent decision support for protein crystal growth. IBM Syst. J. 40, 248–264.

Krupka, H.I., Rupp, B., Segelke, B.W., Lekin, T.P., Wright, D., Wu, H.-C., Tood, P., Azarani, A., 2002. The high-speed Hydra-Plus-

One system for automated high-throughput protein crystallography. Acta Crystallogr. D 58, 1523–1526.
Luft, J.R., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S., DeTitta, G.T., 2001. Macromolecular crystallization in a high throughput laboratory—the search phase. J. Cryst. Growth 232, 591–595.
McPherson, A., 1982. Preparation and Analysis of Protein Crystals. Wiley, New York.
McPherson, A., 1999. Crystallization of Biological Macromolecules. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
McPherson, A., 2001. A comparison of salts for the crystallization of macromolecules. Protein Sci. 10, 414–422.
Norvell, J.C., Zapp-Machalek, A., 2000. Structural genomics programs at the US National Institute of General Medical Sciences. Nat. Struct. Biol. Suppl. 7, 931.
Rupp, B., 2003. High throughput crystallography at an affordable cost: the TB structural genomics crystallization facility. Acc. Chem. Res. (in press). DOI: 10.1021/ar0200216.
Rupp, B., Segelke, B.W., Krupka, H.I., Lekin, T., Schafer, J., Zemla, A., Toppani, D., Snell, G., Earnest, T.E., 2002. The TB Structural Genomics Consortium crystallization facility: towards automation from protein to electron density. Acta Crystallogr. D 58, 1514–1518.
Schick, B., Segelke, B.W., Rupp, B., 2001. Statistical analysis of protein crystallization parameters. ACA Meeting 2001, Los Angeles, CA. American Crystallographic Association Meeting Series 28, p. 166.
Segelke, B.W., 2001. Efficiency analysis of sampling protocols used in protein crystallization screening. J. Cryst. Growth 232, 553–562.
Segelke, B.W., Rupp, B., 1998. Beyond the sparse matrix screen: a web service for randomly generating crystallization experiments. American Crystallographic Association Meeting Series 25, p. 78.
Sivia, D.S., 1996. Data Analysis: A Bayesian Tutorial. Oxford University Press, Oxford.
Stevens, R.C., 2000. Curr. Opin. Struct. Biol. 10, 558–563.
Stura, E.A., Nemerow, G.R., Wilson, I.A., 1992. Strategies in the crystallization of glycoproteins and protein complexes. J. Cryst. Growth 122, 273–285.
Vilasenor, A., Sha, M., Thana, P., Brauwner, M., 2002. Fast drops: a high throughput approach for setting up protein crystal screens. Biotechniques 32, 184–189.
Waldo, G.S., Standish, B.M., Berendzen, J., Terwilliger, T.C., 1999. Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. 17, 691–695.