

Predictive models for protein crystallization

Bernhard Rupp^{a,*} and Junwen Wang^{b,1}

^a *Macromolecular Crystallography and TB Structural Genomics Consortium, University of California, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA*

^b *Center for Biotechnology and Department of Chemistry, Temple University, Philadelphia, PA 19122, USA*

Accepted 24 March 2004

Abstract

Crystallization of proteins is a nontrivial task, and despite the substantial efforts in robotic automation, crystallization screening is still largely based on trial-and-error sampling of a limited subset of suitable reagents and experimental parameters. Funding of high throughput crystallography pilot projects through the NIH Protein Structure Initiative provides the opportunity to collect crystallization data in a comprehensive and statistically valid form. Data mining and machine learning algorithms thus have the potential to deliver predictive models for protein crystallization. However, the underlying complex physical reality of crystallization, combined with a generally ill-defined and sparsely populated sampling space, and inconsistent scoring and annotation make the development of predictive models non-trivial. We discuss the conceptual problems, and review strengths and limitations of current approaches towards crystallization prediction, emphasizing the importance of 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 and standardization should be encouraged, aiming to effectively integrate data mining and machine learning efforts into a comprehensive predictive framework for protein crystallization. Similar experimental design and knowledge discovery strategies should be applied to valid analysis and prediction of protein expression, solubilization, and purification, as well as crystal handling and cryo-protection.

© 2004 Elsevier Inc. All rights reserved.

Keywords: High throughput crystallization; Statistical analysis; Machine learning; Structural genomics; Predictive models

1. Introduction

The initial NIH Protein Structure Initiative, PSI-I [1], underway since fall of 2000, provides significant public funding to nine P50 Structural Genomics Centres. One of the main objectives of these centres is the advancement of high throughput crystallography, including methods for high throughput protein crystallization. As a result of these efforts, large amounts of protein crystallization data will become available as the PSI centres are increasing their production during the last months of their funding. One would assume that the analysis of massive amounts of proteomics and crystallization trial

data engendered by the PSI centres should enable deployment of statistical methods and machine learning to develop predictive algorithms for effective protein crystallization with confidence.

There are already some indications emerging that the process may not be as straightforward as it appears. The sources for the difficulties lie fundamentally in the complex physico-chemical nature of protein crystallization, resulting in non-trivial experimental design issues, which again affect data consistency and data validity, and thus in turn determine the applicability and significance of the data mining algorithms that can be deployed to tackle the problem. We will thus have to explore each of these issues in order to evaluate current and possible future approaches towards crystallization data analysis.

A compounding practical issue is that still only few comprehensive reports have emerged on new crystallization

* Corresponding author. Fax: 1-925-424-3130.

E-mail address: br@llnl.gov (B. Rupp).

¹ Present address: Center for Bioinformatics and Department of Genetics, University of Pennsylvania, PA 19104, USA.

statistics and predictions from the initiatives—which already begin to compete for the next round of PSI-II funding announced in early 2004—and a distinct probability exists that under pressure to produce novel structures (which is the ultimate goal of a PSI Centre), the opportunity to create comprehensive and consistent crystallization databases across the centres 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 rendered the publicly available crystallization databases (Biological Macromolecule Crystallization Database, BMCD [2]; Protein Data Bank, PDB [3]) virtually ineffective for the purpose of rigorous statistical analysis and machine learning, even with significant restructuring and annotation efforts [4,5].

In this review, we will discuss the basic challenges resulting from the complex physico-chemical nature of protein crystallization and how they affect all aspects of experimental design and data generation of crystallization experiments. These preliminaries are important, as experimental design and data validity determine critically (and without mercy) the selection, significance, and value of statistical analysis and machine learning employed to derive predictive frameworks of increasing complexity and specificity. We need to introduce more and better defined prior information about specific proteins or classes to achieve true predictive models of value. In the absence of an ultimate and universal crystallization technique, reliance on probabilistic models is for the foreseeable future our best bet to increase specificity, efficiency, and success rates in crystallization trials.

2. Fundamentals of protein crystallization—the experimental perspective

2.1. Physico-chemical basics of protein crystallization

From a phenomenological viewpoint, crystallization is phase separation in a thermodynamically metastable supersaturated system under the control of kinetic parameters, with the favourable outcome being the formation of a crystal. Fulfilment of thermodynamic criteria only implies that crystallization is *possible*, i.e., it is a *necessary but not sufficient* condition for crystallization. Whether the thermodynamically possible outcome is realized depends on the kinetic parameters controlling the process. While the thermodynamic parameters (classifiable into extensive ones like protein or reagent concentrations, and intensive ones such as temperature or pH) are easily controlled by the experimenter, we have only limited influence on (and knowledge about) the kinetic parameters such as equilibration rates, molecular association, preassembly, nucleation [6], and growth

kinetics. These kinetic parameters are pathway dependent and do change, for example, with crystallization method, and vary even with less obvious factors such as drop size [7]. The fact that kinetic parameters determine the actual outcome of a thermodynamically possible scenario also limits the usefulness of otherwise viable diagnostic tools such as static or dynamic light scattering as indicators for success in predictive modelling, as they are based on thermodynamic excess properties (discussed in [8]).

From a microscopic perspective, crystals are periodic assemblies of macromolecules with few and relatively weak contacts between molecules [9]. Detailed packing analysis of crystal structures revealed that protein–protein crystal contact formation appears to be an essentially stochastic process [10]. Moreover, because current protein structure prediction is not accurate enough nor can protein–solvent interactions be modelled with the necessary precision to pinpoint all contributions to the free energy of crystallization (not to speak of the aforementioned kinetic implications), *ab initio* crystallization prediction for proteins is not feasible. Notwithstanding the progress in designing molecular assemblies by engineering of packing contacts between known structural templates [11], the absence of *ab initio* calculations from sequence and physical principles does require statistical approaches to quantify the likelihood of a certain protein to crystallize under given conditions.

2.2. The experimental setup

In contrast to what one would expect considering the fundamental complexity of the crystallization process described above, the actual setup of the common crystallization experiment is deceptively simple: A weakly buffered protein solution is combined in ratios of order 1:1 with a crystallization cocktail, placed in a more or less closed system, and left alone while approaching equilibrium to the point where kinetics permit. The major variables here are the *protein solution*, the *chemical screening conditions*, the *crystallization method*, and the *scores used to quantify* the experimental outcome. Each group provides sources of uncertainty and difficulty for the experimental design.

Rule 1: Crystallization experiments are not as trivial as they may appear from the fact that basic tasks involve little more than pipetting solutions together and closing the system.

2.2.1. Protein

Varying degrees of information are available for the proteins to be crystallized, and often even what is known is not or inconsistently recorded in the crystallization databases. Protein preparations contain various reagents and additives such as lipids, detergents, or cofactors acquired throughout the course of purification, often

only discovered once the structure is determined [12–14]. Each of these components may play a critical role in the crystallization process.² For the success of classification methods, availability and quality of protein-related information is crucial. In fact, statistical analysis of cloning, protein expression, solubilization, and purification shares many similarities with protein crystallization data mining in terms of coping with ill-defined, poorly sampled, multivariate, and clustered sample space (Section 4). The vast majority of the experimental design problems we face when developing predictive models for protein crystallization (discussed in the general statistical analysis Sections 4 and 5) will hamper data mining of protein production data in a similar manner. Given the increasing importance of exploring multiple constructs, either from orthologs [15] or via protein engineering [16–18], to obtain ‘inherently crystallizable’ proteins [19], the need for comprehensive and valid databases for protein production can barely be overemphasized.

Rule 2: Thou shalt record all thy cloning, expression, solubilization, and purification details in *consistent, data mineable* format.

Furthermore, we need to recognize that our protein sample sets are inherently clustered. HTPX efforts largely aim for small prokaryotically expressed, secreted, and highly soluble proteins (a.k.a. low hanging fruit) to meet throughput goals. Such proteins will likely have a different success space and higher success rate expectations than large complex assemblies or membrane proteins [20,21], or eukaryotically expressed proteins, in particular in the presence of conformationally heterogeneous posttranslational modifications and decorations. Remarkable differences in success rates, even for prokaryotic organisms from which the protein samples were derived, have already been reported [22].

2.2.2. Chemical parameter basis set

Given the practically unlimited number of combinations of chemical components in crystallization recipes, it comes as no surprise that crystallization conditions were empirically 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 [23], and abundant variations of this first kits are now available (although, as noted before [24], the rationale for their design is sometimes less than crystal clear). Most crystallization screen designs pre-classify reagents into one or more groups such as precipitant, additive, buffer, detergent, etc. [23,25], and contain combinations of one (or none) reagent out of each of these classes. In a statistical sense, repeated use of such premixed ‘sparse matrix’ solutions amounts

to oversampling of certain spots in the multidimensional crystallization space. Reported success rates thus are limited to few specific combinations of a pre-selected basis set of reagents. Nonetheless, these screens are successfully used even in high throughput efforts [15] and allow straightforward basic success rate analysis [22,26].

While relative virtues of chemicals comprising a basis set can be readily analysed (Section 6), method-determined variables such as drop size and protein concentration can be as influential and valuable [27] parameters as the reagents used, the systematic investigation of non-chemical parameters spanning the crystallization space has received less attention [28]. Regardless of what crystallization design is used, non-overlap in crucial parameters, be it reagents or technical parameters, does make unbiased comparison of results between different research groups (or designs) rather difficult.

Rule 3: Be aware that there are probably more parameters in a crystallization experiment which you cannot control than the ones you can.

2.2.3. Crystallization setup techniques

Numerous crystallization techniques have been developed, each with different merits and drawbacks depending on the specific purpose (initial screening, optimization, and harvesting) and type of research environment [8]. Fig. 1 provides a quick overview of a few popular methods and some of their (dis)advantages. Many more special techniques are described in [29].

Maximizing comprehensive sampling of crystallization space with the least amount of material plus ease of miniaturization favours robotic high throughput setup of nano-drops [30,31]. In addition, absence of sealing requirements makes micro batch methods under oil [32–35], and free interface diffusion micro-crystallization in multi-layer soft lithography chips [36,37] attractive. Excluding proteins which are unlikely to crystallize from expensive expression and purification scale-up via micro-crystallization techniques used in combination with integrated micro-purification [38] could provide a significant efficiency gain—despite possible loss of some proteins due to non-transferability to a different optimization and harvesting technique (Fig. 2).

Optimization is generally pursued after the major parameters (or factors, compare experimental design Section 4) have been determined during screening. We estimate that about 10–20% of prokaryotic proteins may diffract well out of initial screens. This metric is seldom reported, but astonishing values up to 86% have been observed [41] in nanodrop vapour diffusion setups. A systematic study indicated higher success rates of micro-batch screening in direct comparison with vapour diffusion experiments [40]. From a process efficiency viewpoint, the use of the same equipment throughout the process can favour deployment of a single technique for

² More succinctly phrased, ‘Proteins in solution are sticky, dirty creatures that pick up no end of detritus from their environment’ [29].

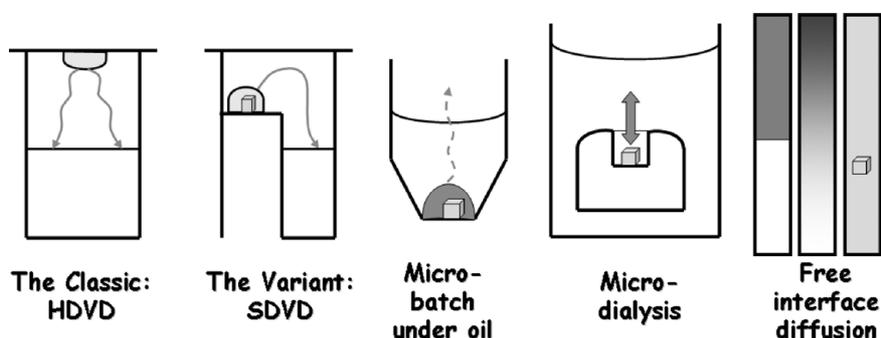


Fig. 1. Schematics of popular crystallization techniques. Hanging drop vapour diffusion (HDVD) is the most popular method in manual setups, and sitting drop vapour diffusion (SDVD) is commonly used with robotic nanodrop setups. Absence of additional sealing requirements and ease of miniaturization favours automated microbatch screening under oil, although harvesting tends to be more difficult. Microdialysis is difficult to automate or to miniaturize. Miniaturized free interface diffusion screening chips are gaining popularity, but automation and harvesting issues remain to be resolved. Each method traverses the crystallization phase diagram in a different path and the *same chemical screening conditions* do not necessarily produce the same results.

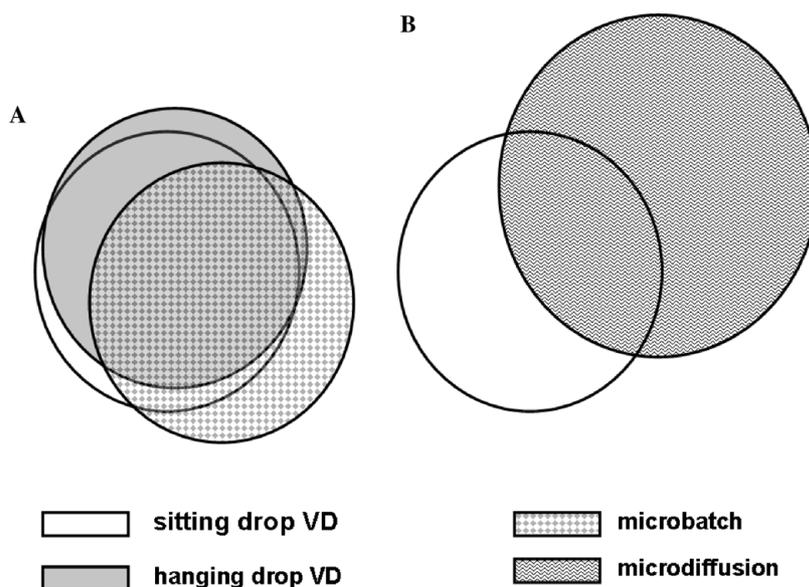


Fig. 2. Venn diagrams representing crystallization scenarios, reproduced with permission from [39]. (A) Overlap in success space between three different techniques. Circles have the same diameter, indicating equal overall success rate for each method. Overlap between hanging and sitting drop vapour diffusion (VD) techniques is presumably large, whereas microbatch may have fewer conditions in common with either [40]. (B) Hypothetical scenario representing free interface diffusion micro-technique with potentially 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 and designs.

screening and optimization [8], enabling consistent robotic set up of either grid-type or limited random-type optimization experiments subsequent to screening. Some not yet confirmed evidence points towards increased screening success rates in TOPAZ (Fluidigm) microfluidics chips (B.W. Segelke, LLNL, personal communication). From a chemical standpoint, the distinction between crystallization screening and optimization of crystal growth is arbitrary, but a change in method related parameters can significantly impact overlap and success rates (Fig. 2).

2.2.4. Scoring of crystallization trials

The majority of crystallization laboratories quantify their crystallization results by assigning quality scores on

a Q -scale with a granularity of 3–10. For example, 0 may indicate a clear drop, and 9 a perfectly looking, isotropically dimensioned crystal. A Q -scale of 1–7 may be used [28], or only three levels, clear, precipitate, crystalline, may be assigned [22]. With proper training and cross-validation, individuals as well as automated image recognition and scoring routines [42–45] can assign scores in a relatively consistent way *within* a given laboratory. Unfortunately, what the scores really mean varies, and no strong common scoring metric exist.

The problems arising from inconsistent scoring are evident. First, in binary models, the cut-off for success can be set rather arbitrarily. For example, in our laboratory [8] we accept a quality score > 5 to indicate success, which includes anything from small crystal clusters via

needles to crystals of perfect appearance. Raising or lowering the cut-off score or varying the definition clearly has a dramatic effect on the reported success rates. This makes inter-laboratory comparison of otherwise potentially interesting statistics such as relative success rates for proteins from different organisms [22] difficult. Second, shape and habit of a crystal can be deceiving. Salt crystals can be scored, and well-developed crystals may show poor diffraction while a cluster of 7's might be dissected to yield a well-diffracting microcrystal fragment. One might then argue that diffraction limit would be the true success score of a crystallization experiment, a point made by 'high-output' proponents [46].

Using diffraction limit as a qualifier for crystallization success, however, carries its own limitations. There are multiple handling steps involved subsequent to observation of a crystal. Harvesting, cryo-protection, mounting and annealing, radiation damage, etc., are all known to affect diffraction quality [47]. These parameters are introduced *after* the crystallization process—and, provided they are screened in a suitable experimental design, do contribute to additional mineable data for the total process (compare Fig. 3), using a combined success indicator 'diffraction quality.' Just as properties of each protein construct and the parameters of protein production influence the outcome of the crystallization experiment and thus need to be recorded for data mining purposes (Rule 2), the post-crystallization handling can provide equally important insights informing about the overall process.

Rule 4: Thou shalt record diffraction limit and all thy harvesting, cryo-protection, mounting, and handling details in *consistent, data mineable* format for each crystal.

3. Data structure, data mining, and experimental design—overview

To successfully *data mine* our crystallization data in order to *discover knowledge* using various *statistical learning* techniques for *descriptive* and *predictive* purposes, we need to be aware of the *quality* and the *structure of the data*, and make an informed decision *which statistical learning technique* is appropriate given the kind of information we want to recover [48,49]. Just turning loose some machine learning algorithm on messy crystallization data and expecting to obtain meaningful patterns is wishful thinking. In simple words, what data do we have and what can we do with them?

3.1. The data repository

Following our rules, we have now amassed an enormous amount of data. They have different nature, and play a role and are important in different aspects of the

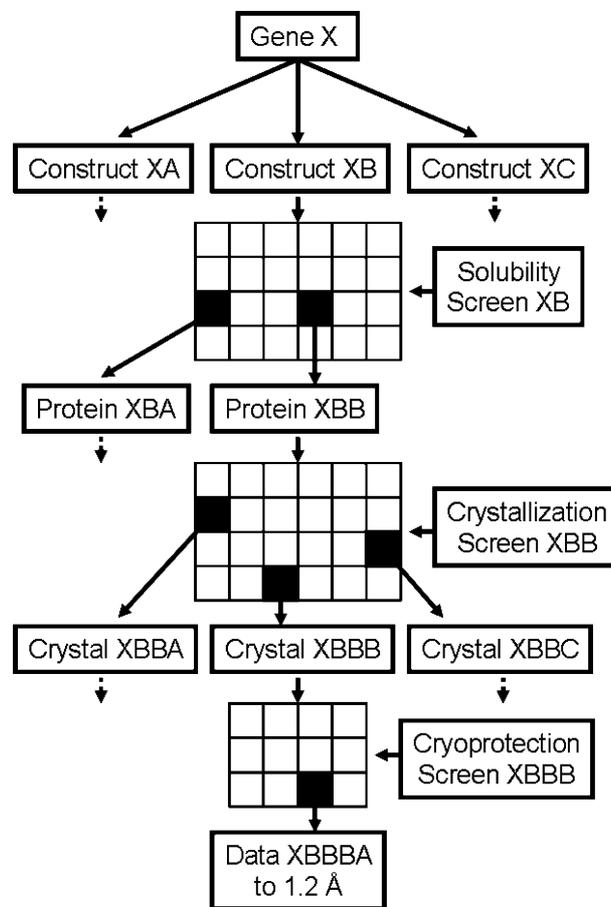


Fig. 3. Path through a hypothetical three-step process tree in a protein crystallization laboratory. In each step invariant parameters from the input can be used for classification of the outcomes of the associated screen. Despite substantial attrition, the amount of data multiplies dangerously [50] with each screening step. The complete set of information from gene to data can be contained as a single instance in case based reasoning techniques [42,48].

analysis. Unfortunately, at this point it may already be decided whether our database is rich with hidden information, ready to be data mined for intelligent decision making support, or whether our crystallization data are presented in a disastrous agglomeration of heterogeneous sources, predestining us as the object of jocular sarcasm by the invigilators of due statistical process. Distributed efforts or consortia are particularly subjected to greatly differing and heterogeneous data repository objects. From the authors' painful experience, the worst legacy types are perhaps spreadsheets scattered across undocumented network file systems.

A good relational database management system will form the core of the data repository, and adequately reflect both the data structure and the process flow [48], and the database design will anticipate the kind of analysis and data mining to be performed [42]. The data repository should also support access to existing databases allowing retrieval of supporting information

that can be used at various levels in the decision making process.

The fundamental physical principles of crystallization and the description of experimental procedures already hint at the major problems we face with our data structure. Our data are incomplete, noisy, and of high dimensionality. While to a certain extent data reduction, cleanup, and normalization (particularly for distance measure and propensity based methods) must be conducted a posteriori, early consideration of suitable experimental design does improve (but not eliminate) most of the source data problems mentioned above. As regards statistical analysis, we can distinguish two basic types of data that we collect. The first major data class are the screening experiment variables and results which we populate our database with. Properties of the material to be screened, on the other hand, remain invariant through each run of an experimental screen and form the basis for classification based on certain properties.

3.2. Screening data and experimental design

When screening a protein for crystallization, we vary the reagents and their concentration in each drop (a treatment) of the experimental run, and record the outcome for each treatment. The experiment could be any type of screening experiment, and with many levels of successively branched screening steps, we follow a decision tree for our process that, despite considerable attrition, creates an enormous amount of data (Fig. 3) that rapidly can outstrip our capability to analyse them [50]. However, we can avoid production of unnecessary or useless data through reduction of *dimensionality* by correlation, principal component analysis, or regression methods, and reduction of *numerosity* by clustering or parametric models [49]. To accomplish this effectively, we need to choose a suitable *experimental design* that is *efficient* and populates the experimental parameter space with sufficiently *complete* and *valid* data for the planned analysis. To appreciate proper design of experiments (DoE), it will be helpful to examine crystallization as a statistical sampling problem.

Rule 5: Bad experimental design resulting in poor data cannot be overcome by whatever degree of posterior sophistication in the analysis. Data quality ultimately—and without mercy—determines the validity of any analysis and the resulting inferences.

3.3. Classification and clustering

Known properties of the material to be screened remain invariant through each run of an experimental screen and form the basis for classification based on certain (class or cluster) properties. For example, the protein has a certain sequence, from which we can derive

information such as *pI*, monomer MW, or presence of an N-terminal His-tag. The protein can have a known biological function, leading to functional classification, or the protein may belong to a certain fold family [51]. Each such cluster may map to a different region in crystallization success space, and using this prior information, one can design a more successful crystallization screen than by following just global crystallization success distributions derived for all proteins. This is, for example, the rationale for all classification or clustering efforts of the BMCD [4,5,52,53].

The distinction between *classification* methods and *clustering* methods is that in the first case we assign a specific class label to each member of the training set and derive a classification rule. A new member of the test set then can be *classified* (predicted) as belonging to a class with a certain probability, based on the ‘learned’ classification rules. As we define or provide the class label, this type of learning is *supervised*. In contrast, we may neither know how many clusters the data contain nor what their class label is. The objective of unsupervised learning is then, based on object attributes, to group objects so that they share similarities within each *cluster*, but are dissimilar (distant) to the objects in other clusters. This grouping is called *clustering*, and the similarity within clusters allows us to treat its objects as one group. Fig. 4 illustrates simple global mapping to hot spots in crystallization space, and Fig. 5 the mapping of a clustered protein sample. *Unsupervised* learning thus discovers additional or novel clusters, and selecting and tuning the proper classification and prediction algorithm is not trivial [49]. Some of our protein classifiers, function for example, may in fact themselves be derived by clustering or combined unsupervised learning techniques [54,55].

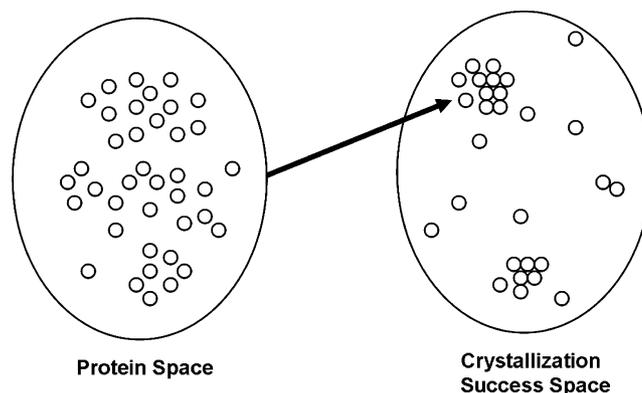


Fig. 4. Global mapping of protein space to crystallization space. The crystallization space has been explored, and two ‘hot spots’ identified, but no explicit classification of proteins has been attempted. Such could represent results from simple crystallization propensity analysis, and the selected hot spot could be a cluster of PEG-MMEs at about neutral pH.

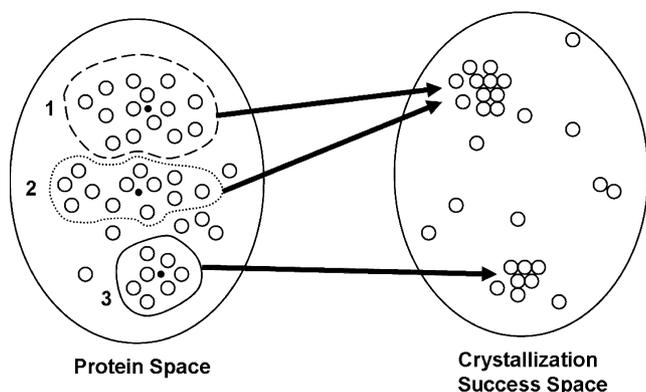


Fig. 5. Mapping of protein clusters to different hot spots in crystallization space. The centroid of each cluster is given by the small full circles. Cluster 3 is clearly distinct from 1 and 2, and also maps to a different region in success space. Clusters 1 and 2 appear less distinct and the fact that they map to the same region in success space indicates that interclass distances may not be sufficient to justify separate treatment.

Commonly used classification techniques are decision tree induction, Bayesian classification and belief networks, artificial neural networks, genetic algorithms, or case based reasoning (some of which are discussed in Section 6 when used for crystallization prediction). Others, like k -nearest neighbour methods, fall prey to the curse of dimensionality, and regression models are briefly discussed in the section on optimization. A full discussion of the concepts and techniques of data mining and machine learning is outside the scope of this review, and we recommend the excellent introduction by Han and Kamber [48].

4. Crystallization screening as a sampling problem

Let us conceptualize the crystallization data space as an n -dimensional vortex or cuboid, whose basis (axes) are extensive parameters like chemical components, protein concentration, and intensive ones like temperature, pH, protein properties, or various setup parameters. Crystallization success analysis can then be treated as a sampling problem of an unknown distribution of successes in crystallization (parameter) space (Fig. 6). Although this picture is simple to understand, the high dimensionality leading to sparse distribution of data points and the varying degrees of freedom (parameters) that can be investigated require careful attention to experimental design.

4.1. Sampling in high dimensional space

Assume a sampling space of dimensionality n . An exhaustive screening experiment can be conducted that varies each dimension n in k steps or levels. The number of experiments to be set up is then k^n . Using only 10 reagents which we systematically vary in three concentration levels (none, low, and high), we need to set up 59,049 experiments. Even using small drops containing only 200 nl of protein solution, we would consume about 12 ml of pure protein preparation, which in all likelihood will meet terminal resistance from our colleagues preparing the material.

4.1.1. Grid screens

Systematic exploration of parameters is feasible only in low dimensions and with low granularity, and a

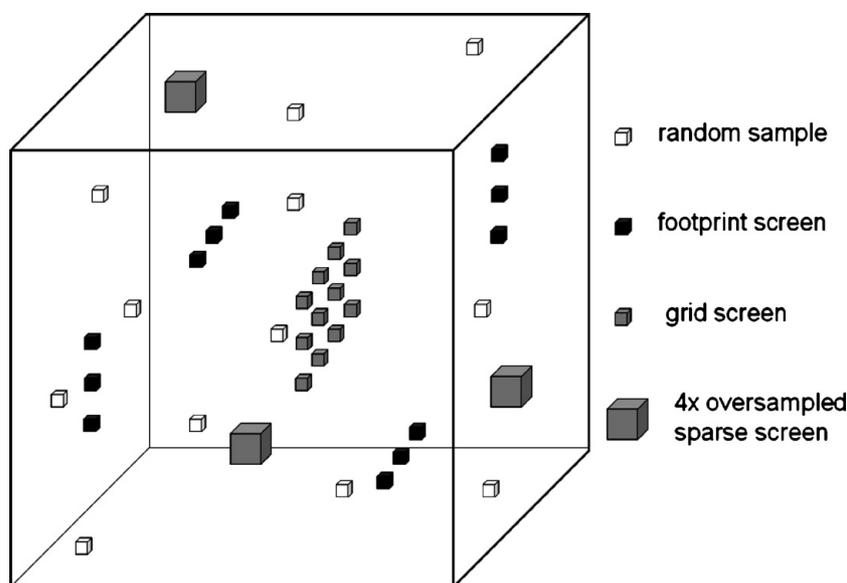


Fig. 6. Schematic of simple 3-d (3 factor) crystallization space showing varying coverage by different sampling protocols using 12 trials each. The large cube represents fourfold oversampling in a sparse matrix-type experiment. Grid screening is not normally used to comprehensively screen for conditions, but deployed with a rationale to explore systematic variations of two dimensions considered major factors while keeping other parameters constant [25]. Optimization with successively finer grids also follows this thought [56]. Representation and grouping of data in n -dimensional data cubes is common practice in multidimensional data mining [48].

typical exhaustive experiment of dimensionality two (varying pH and PEG in four steps) needs only 16 experiments. If the parameters selected (either by intuition or prior knowledge) are indeed dominating factors for crystallization, such designs can rapidly yield valuable information. Initially introduced in a 6×4 format, these designs are known as grid screen experiments [25] and commercial kits are available. Without experimenter bias towards established factors, however, repeated 2-d grid screening and its 1-d variant footprint screening [57] become rapidly inefficient. Given that there are about 400 reagents listed in the BMCD [2,29], a method to effectively distinguish significant factors from nuisance factors that do not contribute to successful outcomes is needed.

4.2. Factorial designs

In view of the impracticality of exhaustive sampling, the need for a rigorous approach towards efficient crystallization screening designs was recognized early by Carter and Carter [60], who suggested factorial experimental designs, which allow application of regression and variance analysis, as well as response surface methods for optimization [28]. Factorial designs attempt to balance the occurrence of probable factors (reagents, pH, and drop size, etc.) and of their combinations during the sampling process. We limit the discussion of designs to the minimum necessary to appreciate its importance for crystallization screening, and refer to Carter [28] and the very readable classical introduction into design of experiments by Box et al. [58].

4.2.1. Full factorial designs

A *full factorial design* is feasible with a limited number of factors and levels of their occurrence. For a two-level (absence and presence), 4-factor (pH, three reagents) full factorial design, 16 experiments are required (24 with replication), and we recognize our 4×4 grid screen experiment (PEG and pH) from the previous section as a complete four-level, 2-factor factorial design. Although a lower number of levels permits to increase the number of factors investigated with the same number of experiments, a complete design becomes rapidly prohibitive. The benefit of the complete factorial is, that similar to a covariance matrix in least squares optimization, it provides a complete picture about all possible interactions between factors [59].

4.2.2. Incomplete factorial design

As the curse of dimensionality limits implementation of full factorial designs, one can reduce the number of experiments by enabling only analysis of first order factor interactions and balancing the design (meaning all factor levels occur with the same frequency) while randomly assigning those factor levels. The resulting design

is an incomplete factorial design [59,60], which can be analysed by means of stepwise multiple regression analysis to identify major factors and their contributions (see optimization section). The key to increasing efficiency a priori in all designs is to select for all *probable* factors, but to *eliminate* irrelevant nuisance factors and *improbable* factor levels (excessive precipitant concentrations, denaturing pH, high PEG concentrations plus high ionic strength [29], etc.) to a priori reduce dimensionality and volume of the n -dimensional sample space vortex. The selection of the major factors PEG and pH for an initial grid screen was thus a well-informed choice [61].

Optimal implementation of incomplete factorials requires specific cocktails and buffers to be prepared for each experiment of a run, and unfortunately, when these systematic statistical designs 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).

4.2.3. Sparse matrix sampling

An approach to reduce the physical effort of setting up experimental designs are ‘sparse matrix’ sampling kits [23]. A basis set of reagents, selected based on prior knowledge about successes (presumed significant factors), is classified into groups such as precipitant, additive, and buffer, and a limited number of non-repeating combinations of one (or none) reagent out of each of these classes are selected. In a statistical sense, repeated use of such premixed ‘sparse matrix’ solutions amounts to oversampling of certain spots in the multidimensional crystallization space. Although success rates thus are limited to relatively few combinations of a pre-selected basis set of reagents resulting in incomplete coverage of the sample space, the original formulations have been successfully used in high throughput screening [22].

4.3. Stochastic sampling

Segelke [19] has assessed grid screen designs [25], footprint designs [57], and sparse matrix designs [23] in terms of sampling efficiency, i.e., finding crystallization conditions with a minimum number of trials. Based on formal statistical derivation, it was demonstrated that random (stochastic) sampling is most efficient, particularly when *success rates are low* or the successes are *clustered*. This 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 ([19], 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 option of protein engineering or search for orthologs should be investi-

gated as viable alternative to wasting resources on continued screening.

In random sampling, coverage of the crystallization space is achieved by using each crystallization condition only once, and robotics are a necessity to practically implement the CRYSTOOL protocol [19,44,62]. Except for the selection of the basis set which also contains various detergents [63], no assumptions about 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 as a serious limitation of random sampling, but there is good reason not to deviate prematurely 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 will be discussed in Section 5.

4.4. Analysis of results for optimization

Although optimization can be considered as a case of crystallization prediction for a specific protein once major factors have been established, optimization analysis is not the subject of this review. Carter [28] provides a concise treatise of linear regression, variance analysis, and response surface methods for the analysis of incomplete factorials and related designs (factorial factorials, orthogonal arrays, and Hardin-Sloane). Neural networks [31] and partial least squares and principal component analysis [64] are also applicable to factorial design optimization. Other multivariate designs used for crystallization optimization are central composite and Box-Behnken design [65] as well as iterative simplex procedures [66].

5. The goal of predictive modelling—a Bayesian viewpoint of protein crystallization

The purpose of crystallization data mining is to *establish relationships* between the experimental parameters and the experimental outcome. The relationships allow deriving *rules and predictions* for the outcome of a new experiment based on the *level of knowledge* we have gained by *prior analysis*. For example, at the lowest level of crystallization data analysis, one seeks to make inferences about global 'hot spots' in success space by basic

frequency statistics, resembling the way the initial sparse matrix set [23] was conceived. At a higher level, one might attempt to categorize proteins into classes with distinct properties, for example, derive special screening kits for membrane proteins [20]. Ultimately, we would like to predict the outcome of a crystallization experiment for a specific protein construct of interest, given all the information we have available.

In other words, important for the long run are correlations (generalizations) of the success distribution in crystallization parameter space with known properties of the protein, i.e., *known priors* that modify our choice of hypothesis (prediction) where to find the highest likelihood for success given what we know already—a classical Bayesian strategy. We thus can classify levels of data analysis according to the level of consideration of prior knowledge (or ignorance) about the problem.

5.1. Bayesian reasoning and inference

A formal basis for defining different levels of complexity can be based on Bayes' theorem, which is derived from the sum and product rules of conditional probabilities [67]. The use and power of Bayesian classifiers and inference is described in many data mining texts [48,68] and we will provide an example of Bayesian classification relevant for crystallization. As used in inference and model testing, Bayes' theorem can be written as

$$\text{prob}(C|D) = \frac{\text{prob}(D|C) \times \text{prob}(C)}{\text{prob}(D)}. \quad (1)$$

Worded as a crystallization example, formula (1) states that the posterior probability $\text{prob}(C|D)$ of our hypothesis (confidence) of crystallization (C) given that the protein belongs to a certain class (D , heavier than 20 kDa and prokaryotic) depends on the probability $\text{prob}(D|C)$ that a sample (protein) of class (D) does in fact crystallize; the prior probability $\text{prob}(C)$ of crystallization *without* consideration of (D)—in our case without considering the protein's class membership; and the probability $\text{prob}(D)$ that the protein is indeed a member of (D). It should be evident that *all* the terms implicitly depend on prior information—none of these terms we know with absolute certainty.⁴ Any prior assumption, made directly or imposed indirectly through a biased experimental design, will thus affect our success prediction. The most striking point is that the posterior $\text{prob}(C|D)$ recovers rather *slowly* from incorrect but plausible priors, whereas a uniform prior (or unreasonable

³ Note, however, that the CRYSTOOL protocol allows adaptive customization of parameter ranges such as pH, reagent concentrations, and reagent frequencies once clearly established trends emerge.

⁴ The explicit conditioning of *all* probabilities on prior information I is often omitted in machine learning texts to simplify algebraic handling. In case of parameter estimation, we can also renormalize the evidence term, formula (1) then is equivalent to $\text{prob}(C|D, I) = \text{prob}(D|C, I) \times \text{prob}(C|I)$ which I have used previously [39].

Table 1
Data objects and their potential use as classification parameters in crystallization data mining

Data object	Examples for invariant properties	Classification examples	Less obvious parameters	Unknown/hidden/ignored
Gene of interest	Sequence, sequence derived information (pI, MW, amino acid distribution, etc.) Function, intrinsic order	Structural families, SCOP, Pfam, COGs [78] Weight, size, functional class	Cellular location, complex partners	Context
Construct	Tags, mutations	N-terminal His C-terminal His	Surface charges Hydrophobicity pattern	Low complexity regions induced by linkers
Protein batch	Purification history	Purity, solubility Monodispersity Additives, decorations	Time between preparation and crystallization Shipping, storage, freezing	Dirt from columns, copurified lipid, cofactor, detergent Residue modifications
Crystallization screens	Instrument parameters	Microbatch Sitting drop VD Microfluidics	Mixing sequence for drops Plate type, surfaces	Kinetics, nucleation Vibration

priors) are more effectively overcome by the likelihood function $prob(D|C)$ (i.e., more or better data). The proportionality in (1) also shows that ‘bootstrapping’ by using the resulting posterior probability as a prior (directly or through premature experimental design changes) and subsequent reanalysis leads to sharpened posterior probability distribution functions, i.e., a serious overestimate of our success rate [67], equivalent to our knowledge model overfitting the data [48]. Both findings are of significance for our crystallization experiments, and caution against early assumptions about success rate distribution in crystallization space, and against a premature limitation of the basis set used in the experiments based on frequency analysis alone. Comprehensiveness thus competes with our desire to reduce dimensionality by sampling focussed on major factors, often at fixed levels (kits) for convenience. Striking this balance is as much influenced by the overall goal of the project (can one accept losses, or has this particular protein to be crystallized regardless of cost?), as it is determined by objective analysis [8]. An interesting predictive effort explicitly utilizing a Bayesian inference model has been described [4], but no recent updates have been made available (<http://www.xtal.pitt.edu/xtal-grow/>).

5.2. Reliability and levels of analysis

The validity of any statistical inference method greatly depends on valid experimental design and the consistency of the databases. In general, we have reasonable control over variable design parameters such as chemical basis set, and over certain (often fixed) parameters of the experimental technique. We have limited or very little control, however, over impurities, batch variation, protein aging, nucleation processes, temperature effects, vibrations, and other ‘hidden’ parameters. Such parameters may be significant factors, and implicitly confound in uncontrolled and irreproducible fashion with other factors. We also obviously cannot make

inferences about parameters we have not explicitly probed.

Similar concerns hold for the quality of our prior knowledge. Poor priors give poor probability distributions. Priors derived from primary sequence like monomer weight, pI, presence and type of affinity tag, etc., are relatively ‘hard’ and reliable. Cellular location, protein interactions, and other prior knowledge that can be relevant for crystallization success may be ‘soft’ and the information itself derived from clustering procedures, and their probabilities less well defined and reliable. As in the case of confounding of factors, uncontrolled parameters such as impurities, co-purified additives or modifications may act as hidden or ‘meta’ priors and can skew the probability distribution of our prior of interest (Table 1).

We can discern, in increasing complexity, three general levels of sophistication in crystallization data analysis. In all cases, both the quality of data *and* the reliability of prior information affect the validity of our hypothesis (prediction) regarding the experimental outcome, $prob(C|D)$. This universal dependence is known as the GIGO principle: Garbage In, Garbage Out.⁵

5.2.1. Analysis without explicit consideration of priors

In the simplest case, we wish to predict what experiments work best to give a desired result by analysing the data using no other *explicit* prior information. The classical case for such an experiment would be global success analysis, simply describing which reagents or conditions work best for all proteins crystallized in a high through-

⁵ This phrase probably originates from a translation of notes by Ada Byron taken during a talk in 1840 by Charles Babbage on properties of his analytical engine [69]. Verbatim she stated: “The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths.” A quotation to keep in mind when attempting to predict protein crystallization using machine learning algorithms.

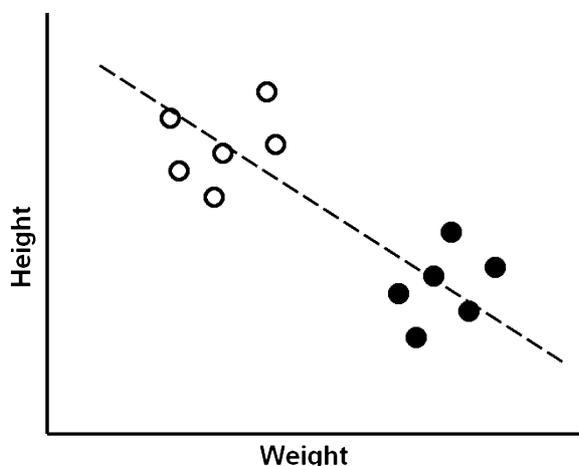


Fig. 7. Sampling of different populations of people (proteins). The left cluster represents the Los Angeles Lakers (rod-shaped proteins), the right cluster Sumo wrestlers (compact and globular proteins). In the case of people, strong prior knowledge prevents us from accepting the hypothesis suggested by the invalid regression, thus: (a) attesting to the power of Bayesian reasoning and (b) demonstrating the adverse effects of neglecting the clustered nature of data. In the example of rod-like vs. globular proteins, there may be considerable less warning signs that a correlation of two given parameters is meaningless.

put crystallization facility. As the use of the term ‘explicit’ prior already indicates, even in this conceptually simple case, experimental design bias (oversampling of limited and sparsely populated parameter space) can compromise the data. Even in random sampling, the selection of the basis set imposes prior assumption on the data—although to a much lesser degree than oversampling with fixed kits. The selection of protein clusters (for example, highly soluble, small, prokaryotic proteins) introduces a strong implicit use of prior information limiting an extrapolation of general validity of the analysis for all classes of proteins. Fig. 7 shows an extreme example illustrating the hazards of establishing correlations based on clustered data.

5.2.2. Analysis using explicit prior information

At the next level of analysis, we might want to predict what experiment works best to give a desired result including prior information about a specific property derived from global analysis of all proteins crystallized before. An example is the use of the protein isoelectric point (pI) as a predictor for crystallization success provided in Section 6. We are using a property of our protein to select a proper prior probability distribution derived from (global) analysis of all proteins. We need to concern ourselves now in addition to the quality of data with the validity of prior information and assure that also our classification prior $prob(D)$ in Eq. (1) is reasonably ‘hard’ and unbiased.

At the most detailed level, we want to predict which experiment has the highest likelihood leading to the

desired result given prior information for a subset or cluster of proteins. In this case, the prior information is derived from a much smaller sample set compared to global priors, and the analysis to establish $prob(C|D)$ is itself more complex and depends critically on the quality of the classification or clustering [49]. Technically we are facing a classification problem in a poorly sampled space of high dimensionality, with all the associated difficulties [48,70]. For such reasons, results from undirected cluster analysis of the BMCD did not translate directly into useful crystallization strategies [5].

6. Practical approaches to crystallization data analysis and predictive modelling

Before we examine practical examples of attempts at crystallization prediction, we must develop a sense of how to assess the quality of the predictions made. These measures are not as straightforward as the basic and familiar moment-based (normal) distribution statistics such as means, variation, and p values. Too many patterns, for example, reduce the number of instances in each class and thus the significance of associations, and only a pattern that is *interesting* creates also *knowledge*. This is probably easiest demonstrated on hand of interestingness measures for association rules [48]: What is the *support* (what percentage of data objects in fact are covered by the pattern) and how high is the *confidence* (how significant or certain is the association). Given the complexity of crystallization, we have to weigh carefully where to set (subjective) cut-off levels of either descriptors to separate noise from useful associations and predictions.

6.1. Global success space analysis

Using crystallization data from all proteins processed, without any explicit classification, we can analyse the distribution of successful outcomes. For sampling designs using fixed reagent cocktails for each protein, the analysis is straightforward and delivers crystallization success rates for specific conditions and for classes of reagents [22,26]. The results lead to the conclusion that the conditions in the original spares matrix kit [23] are clustered in success space; thus, a smaller subset (reduced dimensionality) can crystallize a significant number of proteins [22]. This fact has been recognized already by the designers of the screen (J. Jancarik, personal communication, 1994) in the form of a 18-condition mini-screen, whose best conditions are within the set of 12 best defined by Kimber et al. Redundancy of conditions appears to be generally high within and between prefabricated screens [41].

In experimental designs where the occurrence of each reagent and combination is neither fixed in frequency nor at levels, and where not each protein receives the same treatment, normalization must occur and clustering of conditions is more subtle and difficult to visualize. Such is the case for random screening, and we use the normalized statistic of crystallization *propensity* to quantify the usefulness of a reagent.

6.1.1. Crystallization propensity analysis

Propensity analysis, also called normalized frequency analysis, is widely known from its use in protein sequence pattern exploration. Propensity is a measure that assesses whether the frequency of an attribute is more abundant in a subset (propensity >1) or less abundant (propensity <1) compared to the entire set. Crystallization propensity (CP) of a reagent is then defined as the reagent's rate of success (RS) (occurrence of successful trials containing this reagent/occurrence in all trials containing this reagent) normalized by the average success rate (AS) (all successful trials/all trials): $CP = RS/AS$. As propensity is normalized by the average success rate, a propensity of 2 means that the particular reagent is twice

as effective as the average reagent in producing crystals, which is easy to understand.

The simplest question that thus can be answered by propensity analysis in a quantitative way is: which reagents in our basis set are efficient crystallizers (i.e., which have a significantly higher propensity than others) and which are significantly less effective? Such knowledge allows us to optimize the basis set by retaining good crystallizers and to replace weak ones with new reagents without increasing the dimensionality of the sampling space (albeit ignoring the possibility of synergistic reagent effects). We show in Fig. 8 a limited analysis of 230,000 random crystallization trials [19,62] using a basis set of 55 reagents from the TB structural genomics consortium [38,44]. The protein sampling is implicitly biased towards small, highly soluble, prokaryotic proteins, from *Mycobacterium tuberculosis*. Quantile–Quantile plots [71] of our dataset against normally distributed data (not shown) indicate that individual crystallization propensities are distributed near normal, and we can assign standard deviations for each propensity and quantify whether a reagent is a ‘supercrystallizer’ via conventional *p* values.

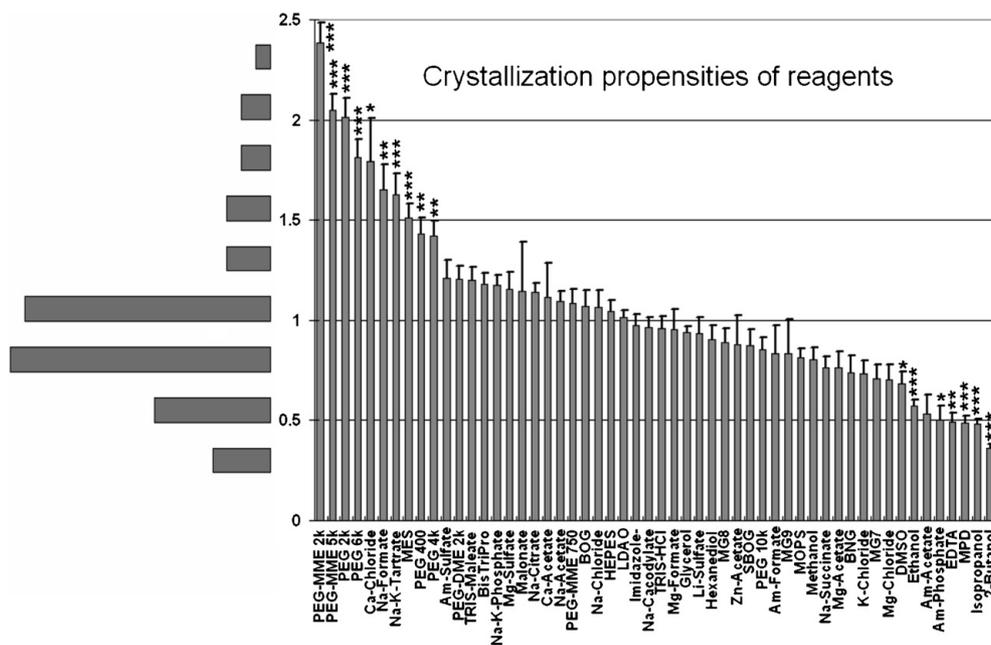


Fig. 8. Crystallization propensity of 55 reagents in random crystallization screening [19], <http://www.doe-mbi.ucla.edu/TB/DB/XTAL/llnl/>. Analysis shows that according to the propensity distribution, PEG-MMEs [72] and PEGs [61] indeed exhibit significantly higher crystallization propensities (*p* values at the confidence levels of 0.05, 0.01, and 0.001 are indicated as *, **, and ***, respectively). Na-malonate [73] shows no significantly increased crystallization propensity, but as malonate has been introduced late into the basis set, and few experiments and successes are available, this statement itself has less significance as asserted by the large standard error bar. The much larger number of experiments and proper normalization now also has confirmed that Am-sulfate, absent in a very limited set analysed earlier for successes [39], shows average crystallization propensity. As a whole class, the small MW alcohols including MPD rank significantly lower in propensity, in agreement with nanodrop results [26]. The finding from analysis of the BMCD that MPD is the single most successful agent promoting crystallization of biological macromolecules [74] is not supported, and probably illustrates the effect of lacking normalization due to absence of negatives in the BMCD (from which the information was derived). It is likely that MPD was just used more often, consistent with its popularity as it acts simultaneously as a cryoprotectant. The need for a sufficiently large and properly normalized sample set to make rare events become significant does not bode well for the proliferation of ‘crystallization tips’ which practically never undergo statistical validation.

The analysis of random screening data [19,44] agrees well with the conclusions from other high throughput efforts [15,22,26,41], and confirms the long-known empirical evidence [61] that PEGs as a class, and in particular PEG-MME's [72] around physiological pH (see also Section 6.2) exhibit a significantly higher crystallization propensity. Although we still do not have enough data to comment with confidence on the propensity of malonate, the high propensity for tartrate and citrate [15] supports the rationale for introducing malonate as a new salt [73].

Propensity scores can also be used to provide quantitative information about the readiness of a protein to crystallize (Fig. 9). Following the normalization procedure described above, we estimate that about 30% of average proteins crystallize with higher than average propensity. This assumption mirrors the overall successes reported from all the structural genomics centres (<http://targetdb.org>), with 37% of all purified proteins yielding crystals. For prokaryotic proteins, individually reported crystallization success rates vary between 37 and 69% [22] on a scale of clear, precipitate, and crystalline, and up to 86% in *T. maritima* [41] on a not further specified scale of 'none,' 'not suitable for mounting,' and 'harvestable.' The score of 'harvestable' here yields only 23% success rate (approximately 7–8 on a scale of 10), and how many actually diffracted right out of the screening is unknown, perhaps around 10–20%.

The absence of a common reporting metric from the centres seriously obscures such estimates. Judging again from the target database, on average only another 39% of those crystallized (only 15% of purified proteins) actually yielded a structure.

Other interesting questions which could be addressed using full random data sets are:

Are there any combinations of reagents (or general factor combinations) that cluster and are significantly more effective than others? Such specific 'hot spots' of course can be sampled with correspondingly higher frequency, and redundant conditions can be replaced with one representative from each cluster and conditions with poor overall propensity (dead spots) could be excluded. Compared to single reagent propensity analysis, they also reveal synergistic effects (factor interactions) between reagents. Given enough data, those synergistic effects may turn out be different for specific classes or clusters of proteins.

Do extreme reagent concentrations (factor levels) exist that yield much less or no successes? The volume of the sample space then can be reduced by truncation.

Are there reagents which crystallize the same proteins equally well (i.e., are linear dependent)? Removal of one of the reagents then reduces dimensionality or allows a new factor to be introduced to the design without increasing dimensionality. Here again is the

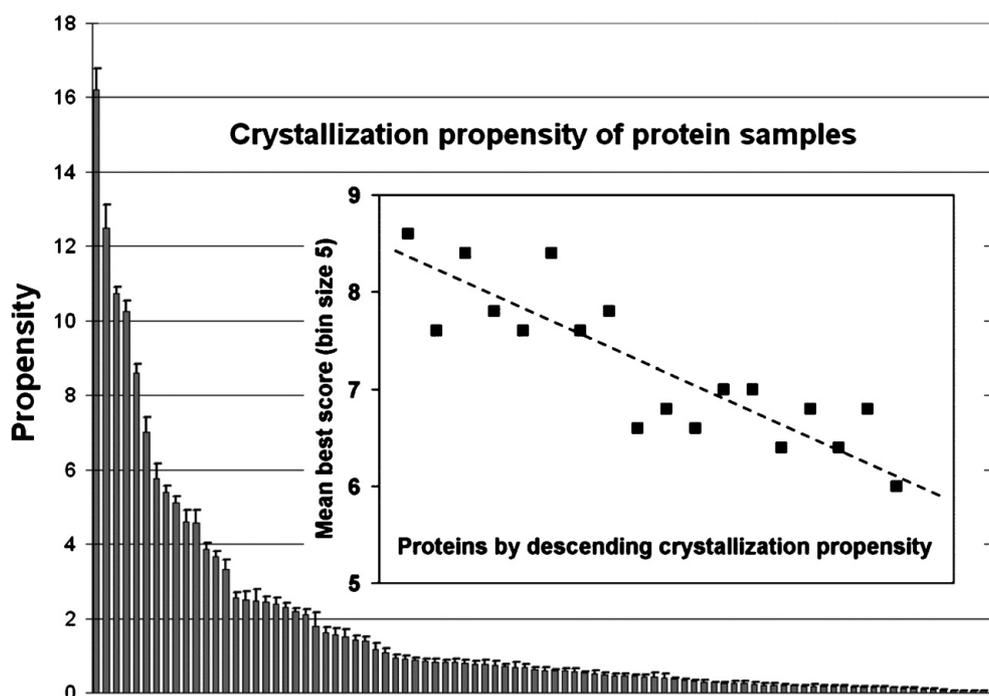


Fig. 9. Crystallization propensity of 91 TB proteins showing at least one success score of 6 or above (<http://www.doe-moi.ucla.edu/TB/DB/XTAL/lnl/>). Correcting for the non-crystallizing proteins, roughly 30% of the proteins crystallize better than average (propensity > 1), which corresponds to the initial estimates [19] that gave rise to selecting 288 as a reasonable number of experiments, beyond which the return on screening investment rapidly diminishes [8,19]. About 150 trials on average are needed per successful outcome. The inset shows that the frequent crystallizers also exhibit higher maximum crystallization scores. The dashed line serves only as a guide to the eye.

need to be able to cluster the random conditions based on the propensities of a large protein sample set.

6.1.2. Outlier detection using market basket analysis

Market basket analysis (MBA) detects patterns of common occurrence of attributes in datasets, and in our case we can search for patterns of reagents occurring together in crystallization successes. MBA essentially answers the question ‘what goes with what’ and the results are presented in the form of clear, readable association rules, whose meaning (confidence and support) is obvious [68]. As we cannot examine an empty basket, MBA can only work on positive results, meaning that evenly distributed samples are needed as we lack a basis for normalization. Although this lack of normalization makes its direct use for discovery of hot spots less reliable than clustering or propensity analysis, MBA has demonstrated its value in outlier detection. Applied to raw success data, MBA eliminated an error in data annotation [39], when we discovered that Na-formate and DMSO formed a binary group in 12% of the cases (support) with an 86% grouping probability (confidence), 5.8 times more likely than random. The result was caused by improper entry of an optimization experiment as a random screen.

6.2. Predictive models using prior information

A practical example for the implementation of what we termed a ‘global’ and reasonably hard prior as an indicator for crystallization success is the use of the protein’s isoelectric point (pI) as a predictor for increased crystallization screening efficiency [75]. The pH of crystallization is distributed monomodally around a mean value of 6.5–7.0, which has been established by analysis of crystallization data in the PDB [2,29,75] and qualitatively confirmed by random data (Fig. 10).

As the crystallization pH is one of the few consistently reported values in the PDB (regardless of crystallization strategy employed) and the pI can be calculated from SEQRES records, one can attempt to correlate the reported crystallization pH with the protein’s pI . From the different shape of the protein pI vs. crystallization pH distribution [75] one already expects in agreement with the literature [2,25,53], that no statistically significant simple *direct* correlation does exist [41]. However, following the empirical pH distribution for a given pI range, protein pI can be used as a predictor to select the pH distribution that maximizes crystallization efficiency (<http://www-structure.llnl.gov/cryspred/>). The well known empirical crystallization pH frequency distribution peaking around 7.0–7.5 can be interpreted as the sum of distributions that favour crystallization by few units above or below pI , respectively [75].

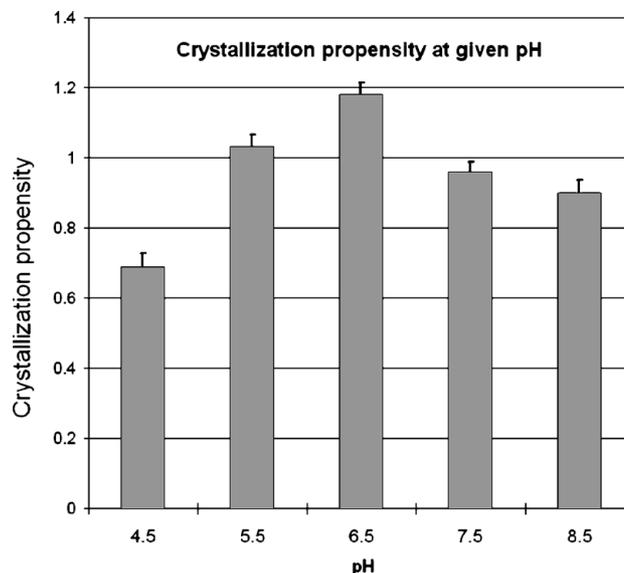


Fig. 10. Crystallization propensity of pH established by random sampling [19,44]. The pH of 6.5 is significantly better than average, and pH 4.5 and 8.5 have significantly less crystallization propensity. Selection of buffers and forbidden combinations (affecting for example phosphate buffers) may lead to under-sampling of extremes of pH, and wider pH range might be screening advisable. Accumulation of more and wider random samples would lead to improved estimators for prototype predictors, which currently may suffer from usage bias in the PDB pH frequency data [75].

Let us investigate a prototype predictor from a Bayesian point of view, according to the terms of equation (1). Our hypothesis is that a protein with a given pI (D) will crystallize most likely at a certain pH, $prob(C|D)$. The quality of this hypothesis then depends on three terms: $prob(D|C)$, the certainty of D conditioned on C , i.e. the probability of pI values (D) being associated with a given pH (obtained from experimental distributions). The term $prob(C)$, the distribution of experimental pH for all crystals, is dependent on how accurately the pHs used to establish the distributions were measured. Similarly, $prob(D)$ describing the distribution of pI for all crystallized proteins, depends on the certainty with which we know the pI . It is again evident, that all the right hand terms depend on the quality of the prior information.

Despite the recognized uncertainty in reported pH, which is rarely measured in the drops, (and nominal values in kits are sometimes deviant [22]), the average error in pH is likely no more than ± 0.5 pH units, and is acceptable. Similarly, we accept that the absolute values of calculated pI s will deviate by about ± 0.5 pH units from the true pI of a protein in solution. Together with necessity for data smoothing and reduction, the confidence limits are compatible with a binning of the teaching set distributions no finer than one pH unit [75]. The part we are least certain of is whether the distribution of the successful crystallization conditions in the PDB is biased and reflects just higher use of certain pH conditions

(buffers).⁶ In our case, the weakest link in (1) thus is $\text{prop}(D|C)$, the likelihood function. With more high throughput screening efforts tracking also negatives, the quality of teaching distributions will increase, generally allowing the implementation of improved Bayesian efficiency estimators [75].

6.3. Data mining and clustering of the BMCD

There have been numerous attempts to extract knowledge from the BMCD [2]. There is no doubt that the BMCD is a valuable resource as far as coarse biasing of reagent basis sets is concerned (evidenced by the success of sparse matrix screens), but the absence of negatives, and thus the number of trials for each treatment, prevents proper normalization and is a serious limitation we have redundantly discussed. Cluster analysis has been employed [5,53] with mixed results, as the clusters are either not significant based on a pseudo- F test, or the rules too general to be of practical value (i.e., have low *interestingness* in machine learning terms).⁷

Based on an augmenting hierarchical classification of the proteins in the BMCD, combined with extensive restructuring, Hennessey et al. [4] have implemented a Bayesian probability calculator in a Windows environment. With a classification basis of about 50 proteins, it would have been interesting to validate the performance of this prior based predictor, but unfortunately, since 2000 there have not been any updates in the literature no on the corresponding web site (<http://www.xtal.pitt.edu/xtalgrow/>).

6.4. Case based reasoning

A very interesting, large scale knowledge based project implementing intelligent decision support for protein crystallization has been built around case based reasoning (CBR) as the machine learning component [42]. CBR has the benefit that it can operate on complex symbolic descriptions as training samples, which allows to treat whole series of processes as cases (compare Fig. 3) as well as to accommodate prior knowledge [48]. Also in this very interesting case we have not been able to find a new update in the recent literature.

⁶ The problem exemplifies the tendency in hypothesis driven science to neglect negative results. Such may work to establish the point (one irreproducibly obtained crystal may suffice to determine a biologically relevant structure), but it cripples discovery based methods by eliminating any confidence in the number of occurrences, which are the basis for strong statistical inference.

⁷ One paper reports the use of Artificial Intelligence [76] to mine the BMCD, with the single but historically unperplexing conclusion that after mid-1983 the use of salts went down while the use of PEGs increased.

6.5. Unexplored machine learning methods for crystallization

Despite the hype surrounding them, artificial neural networks (ANNs) are just a robust tool to approximate complex real-world problems. On the positive side, ANN learning is well suited to problems in which the training data are noisy and complex, and it is one of the most effective learning methods currently known [49]. Similar to decision trees and MBA, ANNs are however not optimal for predominately negative datasets such as crystallization trials. ANNs are also essentially a ‘black box’ system: although we may obtain robust and accurate prediction, the rules (patterns) buried in the system are unknown. A further disadvantage is that ANNs are susceptible to local extrema, which limits the significance of prediction on an absolute scale [49].

To escape the local minima problem, one could deploy genetic algorithms (GAs) as bootstrap techniques. GAs are evolutionary algorithms, a general optimization method that searches a large space of candidate objects for one that performs best according to the fitness functions (selection criteria). GAs generate successor instances by repeatedly mutating and recombining parts of the best currently known instances. These offspring instances are then evaluated by the fitness function and a new population of instances is generated. Although not guaranteed to find an optimal solution, GAs often succeed in finding an object with high fitness [49,77]. Based on their suitability for knowledge discovery in complex and noisy data, we plan to explore the performance of ANNs and GAs together for crystallization data mining of a much more populated random screening database we expect to be available by late 2005.

7. Conclusions

The fundamental questions that arise are: what have we learned so far from the millions of crystallization experiments conducted by the PSI centres, what are the prospects of knowledge discovery, and where are the problems we face now, and how they can be overcome in the future?

We assert that the data mining and machine learning algorithms available today are in principle powerful and plentiful enough to extract useful knowledge from protein crystallization data, thus enabling development of quite sophisticated predictive models for crystallization. Given the high dimensionality, sparse population, and bias inherent in most experimental designs, Bayesian inference models which take into consideration the weakness of our prior information are likely to provide minimally biased estimators for crystallization success. In addition to clear semantics, an added attractiveness of naïve Bayesian inference is its robustness to missing

data, highly relevant to the analysis of sparse (crystallization) screening data. We quote that ‘Over and over, in machine learning people have eventually, after extended struggle, managed to obtain results using sophisticated learning schemes, only to discover years later that simple methods like naïve Bayesian do just as well—or even better’ [68]. Algorithms such as case based reasoning, which allow use of whole decision tree pathways through the complex process of structure determination as elements of training sets, may appear overly ambitious at present. However, like Bayesian inference models, they incorporate prior knowledge and are eminently suitable for grand scale data mining on crystallization databases.

In concord with concerns published recently [50], data warehousing and curating has not received the respect it deserves in view of its enormous importance for successful data mining of highly dimensional and complex proteomics data. If crystallization data analysis is to evolve from basic frequency and propensity analysis to truly predictive models of specific inference, it is mandatory that common metrics for scores and minimal standards for experimental design be mandated. It is also disappointing that no effective, central, and curated repository for the crystallization data from all PSI-I initiatives exists, a situation not likely to change as long as the centres compete with each other for the next round of PSI-II funding. We recommend that in the RFA for PSI-II clearly specified and much more stringent, mandatory requirements for data reporting, metrics, and sharing should be called for.

So far, the crystallization analysis results from the PSI-I initiatives have largely confirmed already known empirical evidence. It is reassuring that minimally biased designs like random screening [19,62] can actually quantify and refine these expectations, and are likely to contribute in the near future the much needed cross-validation of the common, early biased, empirically driven designs.

Specifically, the long-known empirical evidence [61] that PEGs as a class (and in particular PEG-MME’s) around physiological pH exhibit a significantly higher crystallization propensity has been rediscovered by several groups [22,26] and confirmed by unbiased random experiments. The pH preference has been further hardened by classification analysis that provides a rationale for the apparent non-correlation of pI and crystallization pH [75]. An initial, highly efficient primary screen could thus focus around the principal components PEG and pH, supplemented with random variation [19] of other basis set components with high crystallization propensity. The protein sample set itself however is consistently biased towards small, soluble, prokaryotically expressed constructs. At least for this class of proteins, the chemical basis set can be reduced in dimensionality and a random screening strategy

using initially a small subset of principal components with highest crystallization propensity, followed by broader screening if necessary, seems maximally efficient.

It will be informing to see what the final accomplishments in terms of improving crystallization efficiency and the value of predictive models will be at the end of the PSI-I period in fall of 2005, when data mining on a grand scale might become possible.

8. Disclaimer

Transferring a quote from the famous Austrian steam engine designer Karl Gölsdorf (1861–1916) ‘There is no single place on a steam engine where you can save a ton, but 1000 places where you can save a kilogram,’ we feel tempted to state: ‘There is perhaps no single place in protein crystallization (or proteomics) where we can double success rate, but there are many opportunities where we can gain a few percent.’

In the end, no matter how sophisticated the statistical analysis and data mining of crystallization space, any of the techniques will only provide a basis for increasing the *probability* of crystallization success (or in Bayesian terms, increase our *degree of belief* in it), but never *guarantee* success for any particular protein.

Acknowledgments

We thank the current and past members of the TB Structural Genomics Consortium crystallization facility team (B.W. Segelke, H.I. Krupka, B.S. Schick, T. Lakin, J. Schafer, and D. Toppani) for populating the crystallization database. 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, and D. Eisenberg (UCLA); J.C. Sacchettini (Texas A&M University); T. Terwilliger, M. Park, C.-Y. Chang, and G. Waldo (LANL) have supplied a steady flow of proteins used in the crystallization experiments. Li Chen (RCSB Rutgers) has helped in extracting information from the PSI target database. LLNL is operated by University of California for the US DOE under contract W-7405-ENG-48. This work was funded by NIH P50 GM62410 (TB Structural Genomics) centre grant and produced with support of the Reiss Bar, Vienna, Austria.

References

- [1] J.C. Norvell, A. Zapp-Machalek, *Nat. Struct. Biol. Suppl.* 7 (2000) 931.

- [2] G.L. Gilliland, M. Tung, D.M. Blakeslee, J. Ladner, *Acta Crystallogr. D* 50 (1994) 408–413.
- [3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, *Nucleic Acids Res.* 28 (2000) 235–242.
- [4] D. Hennessy, B. Buchanan, D. Subramanian, P.A. Wilkosz, J.M. Rosenberg, *Acta Crystallogr. D* 56 (2000) 817–827.
- [5] R.G.J. Farr, A.L. Perryman, C.T. Samudzi, *J. Crystal Growth* 183 (1998) 653–668.
- [6] J. Drenth, C. Haas, *Acta Crystallogr. D* 54 (1998) 867–872.
- [7] E.R. Bodenstaff, F.J. Hoedemaker, E.M. Kuil, H.P.M. deVrind, J.P. Abrahams, *Acta Crystallogr. D* 59 (2002) 1901–1906.
- [8] B. Rupp, *Acc. Chem. Res.* 36 (2003) 173–181.
- [9] J. Drenth, C. Haas, *J. Crystal Growth* 122 (1992) 107–109.
- [10] O. Carugo, P. Argos, *Protein Sci.* 6 (1997) 2261–2263.
- [11] T.O. Yeates, J.E. Padilla, *Curr. Opin. Struct. Biol.* 12 (2002) 464–470.
- [12] H. Luecke, B. Schobert, H.-T. Richter, J.-P. Cartailler, J.K. Lanyi, *Science* 286 (1999) 255–260.
- [13] M.H. Lamers, A. Perrakis, J.H. Enzlin, H.H. Winterwerp, N. de Wind, T.K. Sixma, *Nature* 407 (2000) 711–717.
- [14] B.P. Klaholz, D. Moras, *Acta Crystallogr. D* 56 (2000) 933–935.
- [15] R. Hui, A. Edwards, *J. Struct. Biol.* 142 (2003) 154–161.
- [16] A.M. Edwards, C.H. Arrowsmith, D. Christendat, A. Dharamsi, J.D. Friesen, J.F. Greenblatt, M. Vedadi, *Nat. Struct. Biol. Suppl.* 7 (2000) 970–972.
- [17] G.S. Waldo, B.M. Standish, J. Berendzen, T.C. Terwilliger, *Nat. Biotechnol.* 17 (1999) 691–695.
- [18] G.E. Dale, C. Oefner, A. D'Arcy, *J. Struct. Biol.* 142 (2003) 88–97.
- [19] B.W. Segelke, *J. Crystal Growth* 232 (2001) 553–562.
- [20] P.J. Loll, *J. Struct. Biol.* 142 (2003) 144–153.
- [21] M.C. Wiener, A.S. Verkman, R.M. Stroud, A.N. van Hoek, *Protein Sci.* 9 (2000) 1407–1409.
- [22] M.S. Kimber, F. Vallee, S. Houston, A. Necakov, T. Skarina, E. Evdokimova, S. Beasley, D. Christendat, A. Savchenko, C.H. Arrowsmith, M. Vedadi, M. Gerstein, A.M. Edwards, *Proteins* 51 (2003) 562–568.
- [23] J. Jancarik, S.-H. Kim, *J. Appl. Cryst.* 24 (1991) 409–411.
- [24] A.M. Brzozowski, J. Walton, *J. Appl. Crystallogr.* 34 (2001) 97–101.
- [25] A. McPherson, *Preparation and Analysis of Protein Crystals*, Krieger Publishing Company, Malabar, FL, 1982.
- [26] D. Hosfield, J. Palan, M. Hilgers, D. Scheibe, D.E. McRee, R.C. Stevens, *J. Struct. Biol.* 142 (2003) 207–217.
- [27] K.V. Dunlop, B. Hazes, *Acta Crystallogr. D* 59 (2003) 1797–1800.
- [28] C.W.J. Carter, *Crystallization of Nucleic Acids and Proteins*, Oxford University Press, New York, NY, 1999.
- [29] A. McPherson, *Crystallization of Biological Macromolecules*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1999.
- [30] B.D. Santarsiero, D.T. Yegian, C.C. Lee, G. Spraggon, J. Gu, D. Scheibe, D.C. Uber, E.W. Cornell, R.A. Nordmeyer, W.F. Kolbe, J. Jin, A.L. Jones, J.M. Jaklevic, P.G. Schultz, R.C. Stevens, *J. Appl. Cryst.* 35 (2002) 278–281.
- [31] L.J. DeLucas, T.L. Bray, L. Nagy, D. McCombs, N. Chernov, D. Hamrick, L. Cosenza, A. Belgovskiy, B. Stoops, A. Chait, *J. Struct. Biol.* 142 (2003) 188–206.
- [32] A. D'Arcy, *Acta Crystallogr. D* 50 (1994) 469–475.
- [33] J.R. Luft, R.J. Collins, N.A. Fehrman, A.M. Lauricella, C.K. Veatch, G.T. DeTitta, *J. Struct. Biol.* 142 (2003) 170–179.
- [34] N.E. Chayen, *Acta Crystallogr. D* 54 (1998) 8–15.
- [35] A. D'Arcy, A. MacSweeney, M. Stihle, A. Haber, *Acta Crystallogr. D* 59 (2003) 396–399.
- [36] C.L. Hansen, E. Skordalakes, J.M. Berger, S.R. Quake, *Proc. Natl. Acad. Sci. USA* 99 (2002) 16531–16536.
- [37] M. van der Woerd, D. Ferree, M. Pusey, *J. Struct. Biol.* 142 (2003) 180–187.
- [38] B. Segelke, J. Schafer, M. Coleman, T. Lakin, D. Toppani, K. Skowronek, K. Kantardjieff, B. Rupp, *J. Struct. Funct. Genomics* 5 (2004) 147–157.
- [39] B. Rupp, *J. Struct. Biol.* 142 (2003) 162–169.
- [40] P. Baldock, V. Mills, P. Stewart, *J. Crystal Growth* 168 (1996) 170–174.
- [41] R. Page, S.K. Grzechnik, J.M. Canaves, G. Spraggon, A. Kreuzsch, P. Kuhn, R.C. Stevens, S.A. Lesley, *Acta Crystallogr. D* 59 (2003) 1028–1037.
- [42] I. Jurisica, P. Rogers, J.I. Glasgow, S. Fortier, J.R. Luft, J.R. Wolfley, M.A. Bianca, D.R. Weeks, G.T. DeTitta, *IBM Systems J.* 402 (2001) 248–264.
- [43] J. Wilson, *Acta Crystallogr. D* 58 (2002) 1907–1914.
- [44] B. Rupp, B.W. Segelke, H.I. Krupka, T.P. Lakin, J. Schafer, A. Zemla, D. Toppani, G. Snell, T.E. Earnest, *Acta Crystallogr. D* 58 (2002) 1514–1518.
- [45] G. Spraggon, S.A. Lesley, A. Kreuzsch, J.P. Priestle, *Acta Crystallogr. D* 58 (2002) 1915–1923.
- [46] N.E. Chayen, *J. Struct. Funct. Genomics* 4 (2003) 115–120.
- [47] E. Garman, *Acta Crystallogr. D* 55 (1999) 1641–1653.
- [48] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.
- [49] T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.
- [50] S.D. Patterson, *Nat. Biotechnol.* 21 (2003) 221–222.
- [51] R. Day, D.A. Beck, R.S. Armen, V. Daggett, *Protein Sci.* 12 (2003) 2150–2160.
- [52] G. Gilliland, D. Bickham, *Methods* 1 (1990) 6–11.
- [53] C.T. Samudzi, M. Fivash, J.M. Rosenberg, *J. Crystal Growth* 123 (1992) 47–58.
- [54] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, D. Eisenberg, *Nature* 402 (1999) 83–86.
- [55] A.H. Liu, A. Califano, *IBM Systems J.* 40 (2001) 379–393.
- [56] M. Cox, P. Weber, *J. Crystal Growth* 90 (1988) 318–324.
- [57] E.A. Stura, G.R. Nemerow, I.A. Wilson, *J. Crystal Growth* 122 (1992) 273–285.
- [58] G.E.P. Box, W.G. Hunter, J.S. Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, Wiley, New York, 1978.
- [59] C.W. Carter, *Methods* 1 (1990) 12–24.
- [60] C.W. Carter Jr., C.W. Carter, *J. Biol. Chem.* 254 (1979) 12219–12226.
- [61] A. McPherson Jr., *J. Biol. Chem.* 251 (1976) 6300–6303.
- [62] B. Segelke, B. Rupp, *ACA Meeting Series* 25 (1998) 78.
- [63] R. Cudney, S. Patel, K. Weisgraber, Y. Newhouse, A. McPherson, *Acta Crystallogr. D* 50 (1994) 414–423.
- [64] J. Zedzik, U. Norinder, *J. Appl. Cryst.* 30 (1997) 502–506.
- [65] P.S. Stewart, P. Baldock, *J. Crystal Growth* 196 (1999) 665–673.
- [66] B. Prater, S. Tuller, L. Wilson, *J. Crystal Growth* 196 (1999) 674–684.
- [67] D.S. Sivia, *Data Analysis—A Bayesian Tutorial*, Oxford University Press, Oxford, UK, 1996.
- [68] I.H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, 1999.
- [69] T. Perl, *Historica Mathematica* 6 (1979) 36–53.
- [70] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, 2001.
- [71] P. Dalgaard, *Introductory Statistics with R*, Springer, New York, NY, 2002.
- [72] A.M. Brzozowski, S.P. Tolley, *Acta Crystallogr. D* 50 (1994) 466–468.
- [73] A. McPherson, *Protein Sci.* 10 (2001) 414–422.
- [74] K. Anand, D. Pal, R. Hilgenfeld, *Acta Crystallogr. D* 58 (2002) 1722–1728.
- [75] Kantardjieff, K., Jamshidian, M., Rupp, B., *Bioinformatics*, 2004, in press.

- [76] A. Rousset, L. Serre, M. Frey, J.-C. Fontecilla-Camps, *J. Crystal Growth* (1990) 405–409.
- [77] K.A. DeLong, W.M. Spears, D.F. Grodon, *Machine Learning* 13 (1993) 161–188.
- [78] C.S. Goh, N. Lan, S.M. Douglas, B. Wu, N. Echols, A. Smith, D. Milburn, G.T. Montelione, H. Zhao, M. Gerstein, *J. Mol. Biol.* 336 (2004) 115–130.