Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Statistical methods for the objective design of screening procedures for macromolecular crystallization

Daniel Hennessy,[a] Bruce Buchanan,[a] Devika Subramanian,[b] Patricia A. Wilkosz[c] and John M. Rosenberg[c]*

[a]Intelligent Systems Laboratory, University of Pittsburgh, Pittsburgh, PA 15260, USA, [b]Computer Science Department, Rice University, Houston, TX, USA, and [c]Departments of Biological Sciences and Crystallography, University of Pittsburgh, Pittsburgh, PA 15260, USA

Correspondence e-mail: jmr@jmr3.xtal.pitt.edu

Received 21 July 1999
Accepted 21 April 2000

The crystallization of a new macromolecule is still very much a trial-and-error process. As is well known, it requires the search of a large parameter space of experimental settings to find the relatively few idiosyncratic conditions that lead to diffraction-quality crystals. Crystallographers have developed a variety of screens to help identify initial crystallization conditions, including those based on systematic grids, incomplete factorial and sparse-matrix approaches. These are somewhat subjectively formulated based on accumulated data from past crystallization experiments. Ideally, one would prefer as objective a procedure as possible; however, that requires objective methods that incorporate a broad source of crystallization data. The Biological Macromolecular Crystallization Database (BMCD), a repository of all published crystallization conditions, is an obvious source of this data. This database has been augmented with a hierarchical classification of the macromolecules contained in the BMCD as well as extensive data on the additives used with them. A statistical analysis of the augmented BMCD shows the existence of significant correlations between families of macromolecules and the experimental conditions under which they crystallize. This in turn leads to a Bayesian technique for determining the probability of success of a set of experimental conditions based on the data in the BMCD as well as facts about a macromolecule known prior to crystallization. This has been incorporated into software that enables users to rank experimental conditions for new macromolecules generated by a dense partial factorial design. Finally, an additional advantage of the software described here is that it also facilitates the accumulation of the data required for improving the accuracy of estimation of the probabilities of success – knowledge of the conditions which lead to failure of crystallization.

1. Introduction

As is well known, successful crystallization is one of the major rate-limiting steps in the determination of a macromolecular structure by X-ray diffraction. During the course of crystallization experiments, substantial data accumulate on the conditions that lead to unsuccessful, partially successful and (hopefully) successful crystallizations. A related problem is coping with the increasing volume of reported results on macromolecular crystallizations. How may these data be exploited in the initial attempts to crystallize a hitherto uncrystallized macromolecule?

The problem of formulating a rational method for macromolecular crystallization has led to many approaches, including the well known incomplete factorial procedure of

Carter & Carter (1979), who assumed that all points in the parameter space of crystallization conditions are equally probable. Jancarik & Kim (1991) employed a semi-automated sparse-matrix sampling of published crystallization conditions to define a crystallization screen that has led to commercially available kits (Hampton Research). Samudzi *et al.* (1992) used simple statistical methods and cluster analysis to postulate qualitative relationships among a subset of crystallization parameters in version 1.0 of the BMCD, leading to a set of suggested screening conditions specific to major classes of macromolecules. In a previous report (Hennessy *et al.*, 1994), we applied machine-learning techniques to the BMCD in order to discover potentially significant new empirical relationships between experimental parameters.

Our approach to utilizing data in published reports of successful crystallizations, such as that summarized in the BMCD (Gilliland, 1987), is to derive probability distributions that form the basis for optimizing search procedures for crystallization conditions for a new macromolecule. One way of stating this is to recall Perutz's famous quote

> crystallization is a little like hunting, requiring knowledge of your prey and a certain low cunning.

Here, the emphasis is on the common impression that patterns of crystallization do exist. Using Perutz's analogy, we ask the following questions. Are there good 'hunting grounds' wherein it is profitable to concentrate one's efforts? Are there poor ones that one should still examine, but not as intensively? Do different classes of macromolecules have different 'favorite haunts' that can be exploited by a cunning crystallographer? Can these haunts be identified *a priori?*

This would not be necessary if one had unlimited resources, one would simply try all possible conditions. In reality, resources are always limiting and most initial screening procedures are intended to approximately locate promising conditions (and obtain solubility data). One concern is that the most promising conditions could 'fall between the cracks' of a coarsely spaced initial screen. If one knew *a priori* that a certain region of parameter space was more likely to yield successful crystallizations than another, one could vary the coarseness of an initial screen so as to concentrate on the more favorable region.

Here, we report results that suggest that different classes of macromolecules indeed show systematic biases in their patterns of successful crystallization. However, before these results could be established, we had to augment the data in the BMCD. In doing so, we limited ourselves to information that would be available before a molecule was crystallized, particularly emphasizing functional information from which structural inferences might be drawn.

The incomplete factorial procedure originally pioneered by Carter & Carter (1979) represents another approach to the crystallization problem. Here too, one is caught between the conflicting demands of the coarseness of the search and the limitations of available resources. One of the unstated assumptions of the original incomplete factorial method is that the macromolecule is equally likely to crystallize at each point in the parameter space of experimental conditions. Here, we report evidence that this is not correct. Even at the outset of a *de novo* crystallization, one can infer from the class of macromolecule to be crystallized that certain regions of the parameter space of crystallization conditions are more likely to lead to a successful crystallization than others. We also describe software that calculates probabilities of success of yielding a diffraction-quality crystal for any combination of crystallization parameters using data retrieval from a modified version of the BMCD. The calculated probabilities are used to bias the selection of data from an incomplete factorial design such that the more probable combinations of experimental conditions are sampled more densely than the less probable ones.

One advantage of the Bayesian approach described here is that it facilitates the development of more objective procedures for the design of crystallization 'screens'. Ideally, one would prefer a totally objective procedure; however, that would require a vast amount of data (essentially every crystallization condition attempted on a broad set of macromolecules, the results of those attempts and considerable data on the physical–chemical properties of the macromolecules). In the absence of those data, Bayesian approaches can still compute ratios of conditional probabilities of success of certain combinations of experimental conditions over others, if we provide them with some prior knowledge about the macromolecules and about the relationships between the different experimental parameters of crystallization. For example, a Bayesian system can conclude from the BMCD data and from prior knowledge about macromolecule families that a 'garden-variety' enzyme has a higher probability of crystallizing at pH of 7.0 than at pH 11.0. As described below, the prior knowledge must be encoded in terms of probabilistic assumptions, which can then be modified on the basis of additional data and experience.

2. Methods

2.1. A macromolecular hierarchy

We have attempted to arrange macromolecules in a taxonomic hiearchy in order to increase the significance of conclusions reached by the program. The motivation for a hierarchical classification was based on several premises. The BMCD contains a variety of molecular types (proteins, DNA, complexes *etc.*) with varying degrees of completeness. Many of these types are not expected to crystallize similarly, *e.g.* proteins and DNA. For others, it was unclear at the outset whether or not there really are meaningful differences; for example, whether or not the heme-containing proteins would crystallize under the same distribution of conditions as 'garden-variety' enzymes. Indeed, one of the fundamental questions we are asking is whether it is even possible to make predictive statements about crystallization conditions knowing only the incomplete information likely to be available at the outset of a crystallization trial.

Table 1
Macromolecular hierarchy.

P	Proteins
P.S	Soluble proteins
P.S.E	Enzymes
P.S.B	Small ligand-binding proteins
P.S.C	Chaperonins
P.S.H	'Heme' proteins (including chlorophyll)
P.S.F	Proteolytic fragments
P.S.I	IgGs and other (soluble) immune proteins
P.S.L	Little proteins and peptides
P.S.L.TI	Toxin/inhibitor
P.S.L.Pep	Short peptides, synthetic or natural
P.S.L.Hor	Hormones
P.S.L.O	Other little proteins
P.S.St	'Structural' proteins
P.S.St.E	Eukaryotic structural (*e.g.* actin)
P.S.St.P	Prokaryotic structural proteins
P.S.St.V	Capsids and other viral components
P.DNA	Proteins that interact with DNA
P.DNA.NS	non-specifically
P.DNA.NS.O	Proteins that bind DNA
P.DNA.NS.E	Enzymes that work on DNA
P.DNA.S	sequence-specifically
P.DNA.S.RM	Restriction/modification enzymes, resolvases, integrases *etc.*
P.DNA.S.AR	Gene activators/repressors
P.RNA	Proteins that interact with RNA
P.RNA.RNP	Proteins from ribonucleoproteins
P.RNA.E	Ribonucleases
P.RNA.T	tRNA synthetases
P.M	Membrane proteins
N	Nucleic acids
N.Z	Ribozymes
N.T	tRNAs
N.O._.n	Oligonucleotides of length *n* (*e.g.* N.O.D.8 is a DNA octamer)
N.O.D.n	Deoxy (DNA) oligonucleotides of length *n*
N.O.R.n (RNA)	Oligonucleotides of length *n*
N.O.H.n	Hybrid (DNA–RNA) oligonucleotides of length *n*
CPP	Protein–protein complex
CDP	Protein–DNA complex
CRP	Protein–RNA complex
CRP.S	'Simple' complexes, *e.g.* tRNA-synthetase
CRP.L	'Large' complexes, *e.g.* ribosomes
CDD	Drug–DNA complex
V	Viruses
S	Sugars and polysaccharides

A closely related issue is the number of entries used to define the probabilities; here, one is caught between the necessity of including enough entries to have a statistically meaningful sample while not defining the system so broadly as to include genuinely dissimilar groups. A hierarchical scheme facilitates adjustment of the sample size. Lower groupings within the hierarchy are more likely to be homogeneous, but they are also smaller. One criterion in developing this hierarchy was that it provide a variety of groupings of different size and possible complexity.

Two important questions to be answered are the following. (i) Can a hierarchical macromolecular classification scheme be found that demonstrates statistically meaningful differences in the distribution of crystallization conditions for its members? (ii) Can these differences be utilized to design crystallization trials? The thrust of this report is to answer yes to both questions. Two additional questions beyond the scope of this paper are as follows. (iii) Is this specific classification scheme optimal? (iv) Can additional *a priori* information be incorporated? We strongly suspect that the answers to these questions are no and yes, respectively; investigation of these questions is central to our long-term research efforts.

We therefore freely admit that the classification scheme shown in Table 1 reflects our own 'world view' of macromolecules. The goal was to achieve a scheme that met the following criteria: (i) it should only use information expected to be available at the outset of a crystallization trial and (ii) it should be based on data that could form the basis of (crude) structural inferences. It should provide classification groups that span a range of sizes from small and specialized to the large and general.

For example, the DNA-binding proteins were segregated because they are expected to have a non-random distribution of electric charge over their surface in order to form stable complexes with highly charged DNA. Similarly, we suspected the structure of heme proteins is influenced by the large prosthetic group; does this also alter their crystallization properties?

The smaller proteins and polypeptides illustrate additional considerations. Here, the proteolytic 'fragments' (P.S.F) were segregated from the 'little' peptides (P.S.L). Size was one consideration in this segregation because the former group consisted of one or more entire domains and were generally larger than the latter those in P.S.L. Size, however, was not the only factor because proteolysis raises the issue of residual protease which could impact on the choice of crystallization conditions. In any event, the software described below allows for the use of multiple categories, *i.e.* the user could combine P.S.F with P.S.L if he/she so chose.

2.2. Restructuring of the BMCD data

To facilitate analysis of the BMCD database, one must convert it to a form more easily amenable to the statistical analysis, machine-learning and the probabilistic screen-design programs. The original BMCD data tables were imported into Microsoft *Access* using its available import capabilities, which allowed the use of SQL[1] queries to augment, partition and analyze the data. Furthermore, this provided a consistent platform by which the *Probabilistic Screen Design* program could access the data using existing well established technology (*i.e.* Microsoft's ODBC[2] tools).

Experimentation and analysis of the data in this format highlighted incompleteness in the data and the need for reengineering including data re-representation, attribute abstraction, data labeling and data subsetting, as explained below. Many of the fields in the BMCD required re-representing or normalization. Symbolic fields (*e.g.* additives or crystallization methods) used multiple terms for the same concept; numeric fields used multiple scales for a single feature. Some attributes represented very complex features

[1] Structured Query Language, an ANSI standard language for manipulating and extracting data in databases.
[2] Open Database Connectivity, a Microsoft standard for interacting with a wide variety of databases.

that were more readily useable as a set of attributes. One example of this attribute expansion was the development of a table relating each of the additives to a set of properties including its perceived role, the number and types of species that comprise the additive and the polymerization state, charge, titrateable type and chemical classification of each species. Furthermore, missing values were handled inconsistently and sometimes overlapped values representing valid entries. The process of re-representing the data to correct for this was important in creating a consistent set of data that would provide maximum utility to later analysis procedures.

Attribute abstraction is the generalization of feature values into hierarchies. Examples of the hierarchies that were developed or exploited include a macromolecular hierarchy (Table 1), a hierarchy of the chemical additives using subsets of the expanded attribute described above and a hierarchy of the space groups. These hierarchies increase the power and flexibility of our analysis by allowing dynamic repartitioning of the data in terms of what is thought to be relevant to the problem at hand; *e.g.* restricting the analysis to immunoglobulin-like proteins when one is attempting to crystallize one.

Data labeling, *i.e.* the identification of positive and negative instances in the data set, was used to analyze the nature of the BMCD data. As mentioned earlier, the BMCD contains only positive instances (*i.e.* successful crystallization results), while most of the machine-learning and statistical analysis procedures benefit from having both positive and negative instances to generate results. The diffraction limit was used to define a sliding scale of 'success' (high resolution) and 'failure' (poor resolution). Thresholds between 2.5 and 3.5 Å were used during various portions of the analysis.

Data subsetting is the selection of portions of the data to reduce the amount of irrelevant or noisy data. This can be of two forms: subsetting of the attributes and subsetting of the database entries. Selection of subsets of the attributes have been used to prune irrelevant or incomplete portions of the BMCD, including attributes which have reported values for fewer than 40% of the entries in the database. (Many BMCD entries are very incomplete, reflecting the original literature on which it is based; *i.e.* in many cases, the actual description of the crystallization conditions lacks many critical variables.) Selection of subsets of the database entries provides more specialized relationships from the data. For instance, the *Probabilistic Screen Design* program focuses the analysis of the probability of success to a specific user-defined subset of the classes of macromolecules, as defined by the macromolecular hierarchy described earlier.

2.3. Statistical analysis of the BMCD

Statistical analysis of the re-engineered data was performed using a two-sample Student's *t*-test to compare the means of the numeric attributes (pH, temperature and macromolecular concentrations) between the different macromolecular classes. The null hypothesis is that the two samples are drawn from the same distribution of values for the attributes. A 'failure' of the *t* test therefore implies that the samples were drawn from two different populations; *i.e.* the means are significantly different. The *t*-test statistic is computed as

$$t = (\overline{x}_1 - \overline{x}_2)/[(s_1^2/n_1) + (s_2^2/n_2)]^{1/2}, \quad (1)$$

where $\overline{x}_i$ is the sample mean, $s_i/n_i^{1/2}$ is the standard error and n_i is the sample size. The Student's *t* test is relatively robust against non-normal populations, especially for larger sample sizes (*i.e.* $n > 15$ given no strong skewness or outliers in the data).

2.4. *Probabilistic Screen Design* – rationale

The minimum number of experiments in a standard incomplete factorial design includes one representative for every pairwise combination of parameter values. Current algorithms can generate more than the minimum number of experiments, *i.e.* more than one representation of every combination. However, they work in an uninformed mode, treating all combinations of parameters as equally likely to produce a successful crystallization. As demonstrated below, all combinations of parameters are not equally likely to succeed. Previous experience can guide the design of the screen to emphasize combinations of parameters with higher likelihoods of success.

A related issue has to do with the coarseness of the sampling. For example, while some proteins crystallize over a broad range of pH, others only crystallize within a very narrow range and promising conditions could fall 'between the cracks' of a sparsely sampled screen. Whenever resources are limited, they impose constraints on the total number of samples to be tested and hence on the 'average' coarseness of any sampling. Our approach is to vary the coarseness of the sampling according to the estimated probability of success, concentrating more closely in the high-probability regions while thinning out those intrinsically less likely to succeed. The degree of variation is under user control and can be easily tuned to the specific problem at hand.

The rationale behind our approach is therefore an extension and combination of the partial factorial (Carter & Carter, 1979) and sparse-matrix (Jancarik & Kim, 1991) approaches. The combinations of conditions in these approaches are based on previous experience (including statistical analysis of previous versions of the BMCD) and anecdotal evidence. However, their relatively *ad hoc* combination of information has resulted in decreased performance, as described below. It also contains a significant subjective component, with all the incumbent risks. In contrast, our approach concentrates much more heavily on analyzing and applying the data in the BMCD as directly and objectively as possible.

2.5. *Probabilistic Screen Design* – probability computation

The data in the BMCD are used to provide an indication of how frequently a combination of specific conditions has produced successful crystallizations in the past. This frequency of occurrence can then be used to compute a relative prob-

ability of success when comparing one set of experimental conditions with another.

The computation of the probability of success is not quite as simple as counting the frequency of occurrence in the database of the given conditions. The BMCD includes many reports of crystals that only diffract to low resolution, *e.g.* 7 Å. Additionally, some types of crystallographic studies require higher resolution than others. For the studies reported here, we adopted an adjustable 'threshold', *e.g.* 2.5 Å, as the minimum required for a 'successful' crystallization. We also assume that the values reported in the BMCD are the most successful combination tested for the macromolecule in question, *i.e.* other unreported values probably yielded crystals that did not diffract as well.

The mathematical formula for the probability that a diffraction limit is under some threshold (or standard of success) given a set of parameters[3] is

$$\begin{aligned} P(D|\mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL}) = \\ [P(D, \mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL})]/ \\ [P(D, \mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL}) \\ + P(\overline{D}, \mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL})]. \end{aligned} \quad (2)$$

Here, $P(D|\mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL})$ represents the probability that the diffraction limit is under the threshold given specified values for the crystallization variables pH, *B* (buffer), *etc*. Similarly, $P(\overline{D}, \mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL})$ is the probability that the diffraction limit is over the threshold.

If we cannot assume any form of independence among the variables, then the only way to evaluate the joint probabilities in (2) is to search the database for the frequency of that exact combination of parameter values. If the database was very large, evenly spanned the full range of conditions and included failure information as well as successes, this technique would do fine. Unfortunately, the BMCD does not fit these criteria.

However, this search for an exact combination of parameter values would only be necessary if we were to assume strong interdependence among *all* of the experimental parameters. As our experience and statistical analysis (see §3.1) both suggest this strong interdependence is not likely, it is reasonable to relax this requirement and allow independence between conditions to be explicitly modeled. This allows us to search for subsets of conditions (something the database can support) and combine them according to the rules of probability. Stated in other terms, the crystallization problem itself would probably be intractable if there were total interdependency of the terms (only one exact combination worked). The observation that multiple combinations give at least partial success is the basis of most crystallization efforts; we are exploiting that here by multiplying the probability functions.

Our model of the parameters and their dependencies is shown in Fig. 1. The advantage of this model becomes apparent in its impact on calculating the joint probabilities in (2). Using the network of dependencies described in Fig. 1, the axioms of probability and Bayes' theorem, the joint probability in the numerator of (2) can be computed as

$$\begin{aligned} P(D|\mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL}) = \\ P(D) * P(\mathrm{MCL}|D) * P(T|D, \mathrm{MCL}) * P(\mathrm{pH}|D, \mathrm{MCL}, T) \\ * P(B|D, \mathrm{MCL}, T, \mathrm{pH}) * P(S|D, B, \mathrm{MCL}, T) \\ * P(\mathrm{SC}|D, S, \mathrm{MCL}, T) * P(\mathrm{PPT}|D, S, \mathrm{MCL}, T) \\ * P(\mathrm{PC}|D, \mathrm{MCL}, \mathrm{PPT}, T) * P(\mathrm{MCN}|D, B, \mathrm{MCL}, T). \end{aligned} \quad (3)$$

The conditional probabilities in (3) can be estimated directly from frequencies in the BCMD. For example, $P(\mathrm{pH} = 7.0|D < \mathrm{threshold}, \mathrm{MCL} = \mathrm{enzyme}, T = 277\ \mathrm{K})$ is the probability that the pH is 7.0 given the diffraction limit is less than the threshold, the macromolecular class is 'enzyme' and the temperature is 277 K. To compute this probability, the number of experiments in the database with a diffraction limit less than the threshold, a macromolecule of the type 'enzyme' and a temperature of 277 K is counted. The number of experiments in a subset of the previous set having a pH of 7.0 is also counted. Dividing the count of the subset by the count of the original set provides the necessary relative frequency and estimate of the conditional probability, shown in (4). Repeating this process for all of the conditional probabilities in (3) and multiplying them results in the computation of the desired joint probability,

$$\begin{aligned} P(\mathrm{pH} = 7.0|D < \mathrm{threshold}, \mathrm{MCL} = \mathrm{enzyme}, T = 277\ \mathrm{K}) \\ = (\text{No. of exps with pH} = 7.0, D < \mathrm{threshold}, \\ \mathrm{MCL} = \mathrm{enzyme}, T = 277\ \mathrm{K})/(\text{No. of exps with} \\ D < \mathrm{threshold}, \mathrm{MCL} = \mathrm{enzyme}, T = 277\ \mathrm{K}). \end{aligned} \quad (4)$$

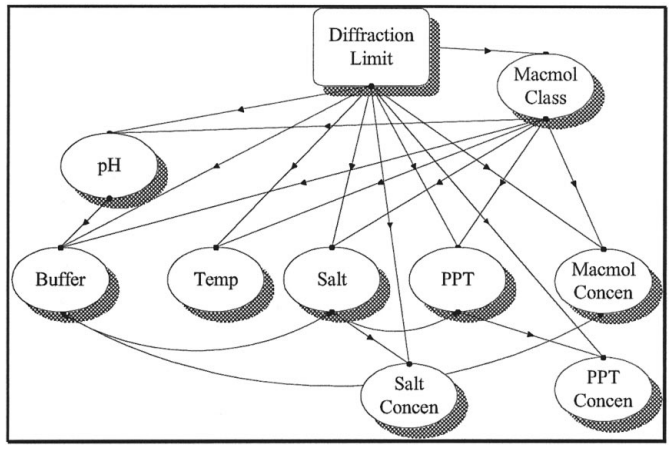

Figure 1
Crystallization parameter dependency graph. The graph represents the parameters included in the calculation of the estimated probability of success and their dependencies. A connecting arc from pH to buffer indicates that the probability distribution for the buffer may depend on the value of the pH. The lack of a connecting arc between two parameters reflects conditional independence (the probability distribution for a parameter is independent of the value of the other parameter).

[3] Parameter legend: *D*, diffraction limit; *B*, buffer; *T*, temperature; SC, salt concentration; PPT, precipitant agent; PC, PPT concentration; MCN, macromolecule concentration; MCL, macromolecule class.

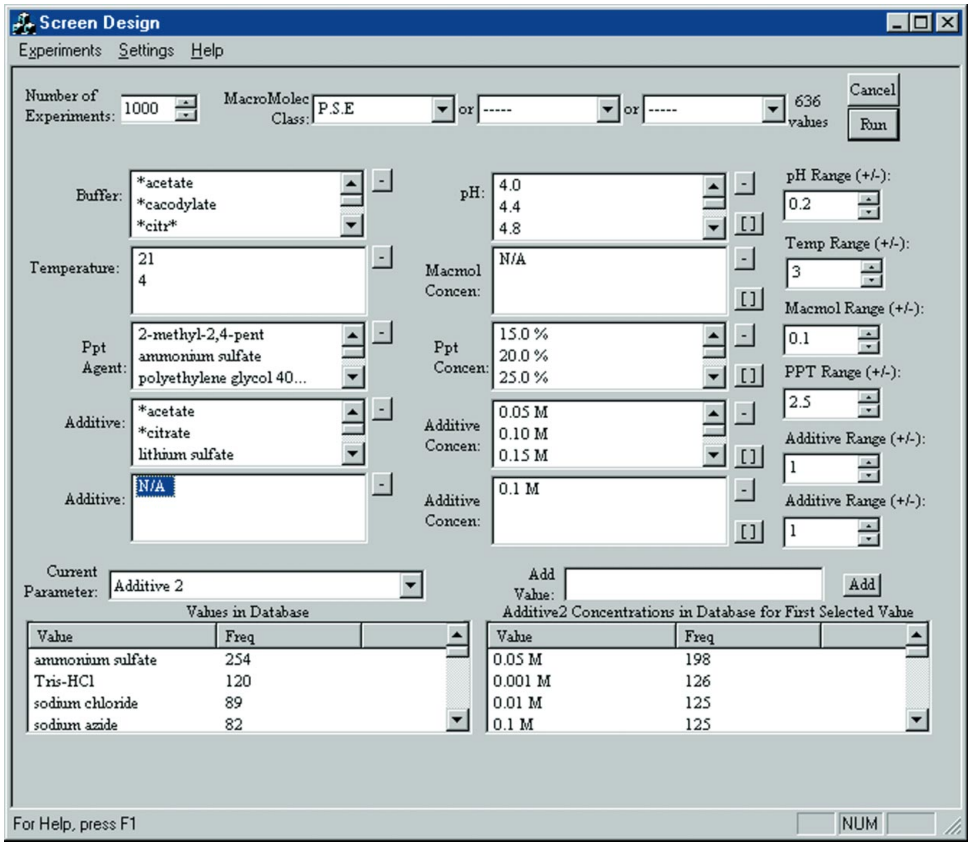

Figure 2
Screen design user interface: a screen dump of the user interface for the *Probabilistic Screen Design* program. The ten large list boxes in the middle of the window are where the values for the crystallization parameters are entered. The two large lists at the bottom of the window display the list of values and their frequencies in the database for a selected parameter. The edit boxes at the top and right side of the window allow for further customization of the screen design. Legend: Temp = temperature, PPT = precipitating agent, Macmol Concen = macromolecule concentration.

Note that these probabilities are dynamic – as additional data is added to the database these conditional probabilities can be updated.

This joint probability in (3) is repeated in (2) as the first operand in the denominator. The second operand in the denominator has an analogous form to (3).

$$P(\overline{D},\mathrm{pH}, B, T, S, \mathrm{SC}, \mathrm{PPT}, \mathrm{PC}, \mathrm{MCN}, \mathrm{MCL}) = P(\overline{D}) * P(\mathrm{MCL}|\overline{D}) * P(T|\overline{D}, \mathrm{MCL}) * P(\mathrm{pH}|\overline{D}, \mathrm{MCL}, T) * P(B|\overline{D}, \mathrm{MCL}, T, \mathrm{pH}) * P(S|\overline{D}, B, \mathrm{MCL}, T) * P(\mathrm{SC}|\overline{D}, S, \mathrm{MCL}, T) * P(\mathrm{PPT}|\overline{D}, S, \mathrm{MCL}, T) * P(\mathrm{PC}|\overline{D}, \mathrm{MCL}, \mathrm{PPT}, T) * P(\mathrm{MCN}|\overline{D}, B, \mathrm{MCL}, T). \quad (5)$$

(5) is problematic in that it requires frequencies for failures of crystallization experiments. This information is not available in the BMCD. Over time, we hope to augment the BMCD with data from additional experimentation that will provide some of this failure data. Until then, we have estimated these by assuming that the shape of the probability distribution for failure is complementary to the shape of the probability distribution for successful crystallizations. For most of the parameters, this seems to be a reasonable assumption.

Therefore, extracting the conditional probabilities in (3) from the BMCD and applying the assumptions and prior probabilities to determine the conditional probabilities for (5), the necessary joint probability distributions can be calculated and substituted into (2). The result is the probability of success given a specific set of parameter values based on the data in the BMCD.

Using (2) through (5), a probability of success can be computed for each of the sets of parameters produced by an algorithm that generates crystallization conditions, such as the Carter & Carter algorithm (Carter & Carter, 1979). Note that although we used the Carter & Carter algorithm in this implementation, there is nothing inherent in our probability calculations that would limit them to this method for generating the initial conditions to be tested. The probability can be used as the basis for a scoring mechanism to bias the selection of experiments that are actually carried out. The result is a screen design, dynamically created based on the most current information, where previous successes indicate the greatest likelihood of future success.

2.6. *Probabilistic Screen Design* – program description

The biased incomplete factorial design algorithm described above has been implemented for Microsoft *Windows* environments.[4]

Once invoked, the program displays help boxes which walk the new user through the setup and execution procedures. Information is gathered from the user describing the set of parameters around which the crystallization screens will be designed. Input typically includes information about the fixed experimental conditions, such as the class of macromolecule, the parameters to be manipulated with the values they will assume and the number of trials. Fig. 2 shows the user interface for the screen-design program.

For the initial version of this program, the set of parameters that defines the crystallization trials is fixed to include pH, temperature, macromolecular concentration, buffer, two

[4] A Java version is currently under development with a direct connection to a database derived from the BMCD.

Table 2
A systematic correlation between the protein family and the temperature distribution (lower left) and the pH (upper right).

This table reports the values of the Student *t* test, as applied to the indicated pair of variables. The value, quoted as a percentage, represents the probability that the two individual distributions were in fact drawn from the same parent distribution. For example, the value of 0.02% in the P.S.E–P.S.B comparison (upper right) means that the probability that both pH distributions were drawn from the same parent distribution is only 2×10^{-4}, which is highly significant. The quantity in parentheses is the value of the *t* function itself.

	P.S.E	P.S.B	P.S.I	P.S.H	P.S.St	P.S.L	P.M	P.DNA	P.RNA
P.S.E		0.02 (3.81)	—	8.1 (−1.74)	—	—	9.5 (−1.73)	5.6 (−1.94)	—
P.S.B	—		6.8 (−1.83)	0.01 (−4.85)	—	4.6 (−2.00)	0.4 (−3.08)	0.01 (−3.95)	—
P.S.I	0.14 (−3.32)	—		8.6 (−1.73)	—	—	5.69 (−1.97)	3.5 (−2.14)	—
P.S.H	0.96 (−2.61)	—	—		9.3 (1.72)	—	—	—	6.2 (1.88)
P.S.St	—	—	9.1 (1.72)	—		—	3.6 (−2.15)	3.6 (−2.14)	—
P.S.L	4.1 (−2.06)	—	—	—	—		9.5 (−1.70)	8.8 (−1.72)	—
P.M	0.01 (−6.84)	0.02 (−4.03)	4.2 (−2.07)	0.02 (−4.02)	0.01 (−3.39)	1.5 (−2.46)		—	4.5 (2.00)
P.DNA	—	—	—	—	—	—	0.70 (2.78)		2.5 (2.20)
P.RNA	—	—	—	—	—	—	0.01 (3.60)	—	

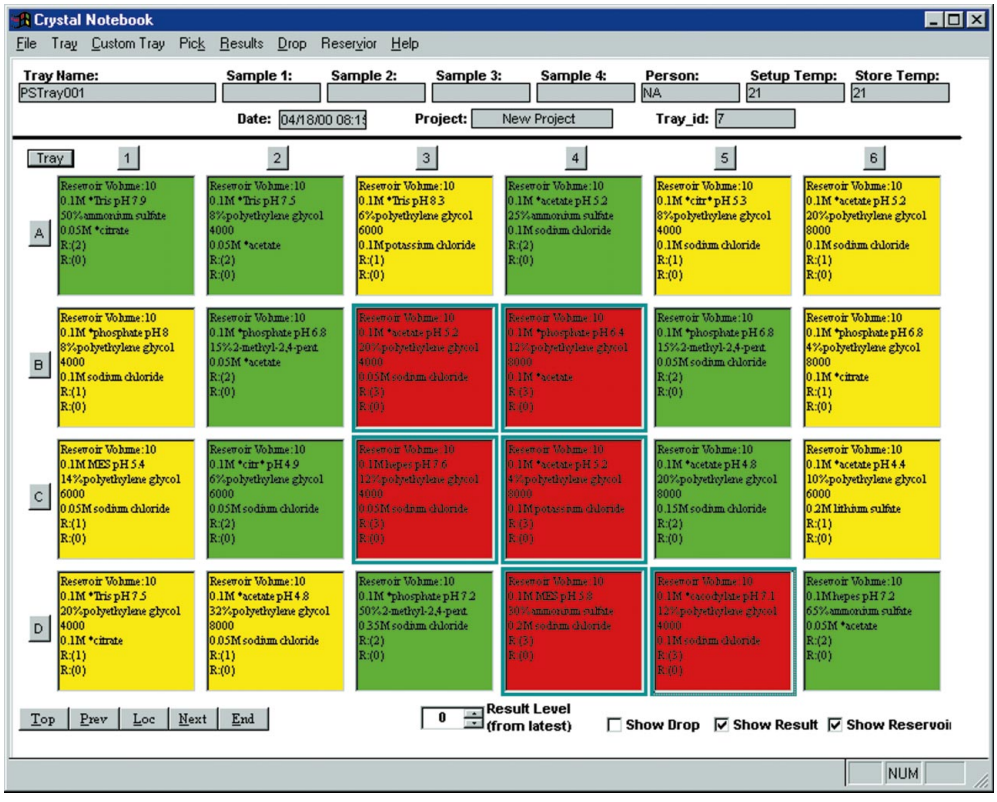

Figure 3
Crystallization Notebook user interface: a screen dump of the user interface for the *Crystallization Notebook* program. The 24 boxes in the middle of the window represent the 24 'wells' of a typical crystallization experiment. Solutions for each well along with the list of results are displayed in each box. User-customized colorization supports a 'phase-diagram' display of results. Menus support flexible entry of recipe/concentration data, mass editing, user-defined 'default' tray setups, a Grid Screen wizard and printed output for inclusion in experiment notebooks.

additives (one of which is generally a salt) and precipitating agent. A listing of potential values for the parameters and the extent to which they are represented in the database is provided at the bottom of the window (see Fig. 2). The parameter values support the use of wildcards ('*' and '?') for matching partially specified names in the database. This is useful, for instance, when the user is interested in the impact of using a 'phosphate' buffer regardless of the anion. Binning ranges are supported for the numeric values to aid in matching numeric values against a database. The binning range specifices how strict a numeric match is required. For instance, a binning range of ±0.2 for the pH allows 7.5 to match any value between 7.3 and 7.7 inclusive.

Given these parameter values, the program selects a set of trials that best covers the space of possible combinations of these crystallization parameters. The user specifies the number of desired trials, which are generated by the standard Carter & Carter algorithm. For each of the generated trials, the necessary frequencies are extracted from the database and the probability of success is computed as described above. The list of experiments, including parameter values, rank order and computed probability of success, is presented to the user for selection and grouping into trays.

2.7. *Crystallization Notebook* – program description

During the course of crystallization experiments, substantial data accumulate on the conditions that lead to unsuccessful, partially successful and (hopefully) successful crystallizations. One of the goals of this project is to provide for the unobtrusive

Table 3
Conditions generated by screen-design program for enzymes.

P.S.E, 294 K

*Tris pH 7.9, 50.0% AS, 0.05 *M* *citrate	*Tris pH 7.5, 8.0% PEG 4000, 0.05 *M* *acetate	*acetate pH 5.2, 25.0% AS, 0.10 *M* NaCl	*citr* pH 5.3, 8.0% PEG 4000, 0.10 *M* NaCl	*acetate pH 5.2, 20.0% PEG 8000, 0.10 *M* NaCl	*phosphate pH 8.0, 8.0% PEG 4000, 0.10 *M* NaCl
*phosphate pH 6.8, 15.0% MPD, 0.05 *M* acetate	*acetate pH 5.2, 20.0% PEG 4000, 0.05 *M* NaCl	*phosphate pH 6.4, 12.0% PEG 8000, 0.10 *M* acetate	MES pH 5.4, 14.0% PEG 6000, 0.05 *M* NaCl	*citr* pH 4.9, 6.0% PEG 6000, 0.05 *M* NaCl	HEPES pH 7.6, 12.0% PEG 4000, 0.05 *M* NaCl
*acetate pH 4.8, 20.0% PEG 8000, 0.15 *M* NaCl	*acetate pH 4.4, 10.0% PEG 6000, 0.20 *M* Li_2SO_4	Tris pH 7.5, 20.0% PEG 4000, 0.10 *M* *citrate	*acetate pH 4.8, 32.0% PEG 8000, 0.05 *M* NaCl	*phosphate pH 7.2, 50.0% MPD, 0.35 *M* NaCl	MES pH 5.8, 30.0% AS, 0.20 *M* NaCl
*cacodylate pH 7.1, 12.0% PEG 4000, 0.10 *M* NaCl	HEPES pH 7.2, 65.0% AS, 0.05 *M* *acetate	*acetate pH 4.8, 8.0% PEG 4000, 0.15 *M* *acetate	*Tris pH 7.5, 12.0% PEG 4000, 0.35 *M* NaCl	MES pH 6.6, 16.0% PEG 8000, 0.05 *M* *acetate	*citr* pH 4.9, 25.0% AS, 0.10 *M* *acetate

P.S.E, 277 K

*Tris pH 8.3, 6.0% PEG 6000, 0.10 *M* KCl	*phosphate pH 6.8, 15.0% MPD, 0.05 *M* NaCl	*phosphate pH 6.8, 4.0% PEG 8000, 0.10 *M* *citrate	*acetate pH 5.2, 4.0% PEG 8000, 0.10 *M* KCl	*Tris pH 7.5, 35.0% AS, 0.10 *M* *acetate	*phosphate pH 7.6, 16.0% PEG 8000, 0.20 *M* *acetate
HEPES pH 7.2, 15.0% MPD, 0.05 *M* *citrate	*citr* pH 5.7, 50.0% AS, 0.05 *M* KCl	*citr* pH 5.7, 8.0% PEG 4000, 0.10 M *citrate	*citr* pH 4.9, 6.0% PEG 6000, 0.05 *M* *acetate	*citr* pH 5.7, 14.0% PEG 6000, 0.05 *M* NaCl	HEPES pH 8.4, 25.0% AS, 0.05 *M* *citrate

recording and archiving of crystallization experiments, including the incorporation of tools for performing chemical and related calculations. Toward this end, the *Crystallization Notebook* database software package has been developed for Microsoft *Windows* environments.

Included in the *Crystallization Notebook* experiment-recording package is a completely integrated graphical user interface for data recording (see Fig. 3). Solutions for each well are displayed in the format of an entire tray for easy readability. Additionally, the software supports user customized and colorized 'phase-diagram' display of results, flexible recipe/concentration data, the ability to select and mass edit any subset of the wells of a tray, user-defined 'default' tray setups, the ability to 'derive' a tray from a cell capturing the natural flow of experiment, a Grid-Screen wizard to provide quick grid-screen generation and entry, the ability to compute concentration data from recipe and the recipe from concentration, and printed output for inclusion in experiment notebook.

Both software packages are available for distribution; users may register and download the software from http://www.xtal.pitt.edu.

3. Results and discussion

3.1. Statistical analysis of the BMCD

Version 2.0 of the BMCD consists of 2353 successful crystallization conditions for over 1557 macromolecules such as proteins, nucleic acids, polysaccharides and viruses (Gilliland, 1987). As many as 53 parameters are reported per crystallization entry, including pH, temperature, method, macromolecular concentration and chemical additives to the growth medium, *e.g.* precipitating agents, buffers and salts. Descriptive information includes the 'common' macromolecular name, the E.C. classification and the source, *e.g.* organism and/or tissue where appropriate. Crystallization results are reported: unit cell/space group, crystal size and shape, X-ray diffraction limit and crystal-growth time.

We expect that most practicing protein crystallographers would agree that the successful crystallographers exploit patterns of crystallization. However, there has been little statistical data to provide objective support for this belief. We therefore applied Student's *t* test to the BMCD, as described in §2. Specifically, we addressed the question: are there meaningful differences between the macromolecular classes in their distributions of temperature, pH *etc*. as reported in the BMCD? The results shown in Table 2 show that this is the case. Here, we show the value of the *t* statistic together with the probability that that value occurred by chance. Probabilities under 1% are generally considered significant and those under 0.1% are highly significant.

As can be seen, there are several highly significant differences in the distribution of temperature and/or pH values for the different macromolecular classes. For example, the distribution of pH values in the BMCD for the ligand-binding proteins (P.S.B) is significantly different from those reported for enzymes (P.S.E) at the 0.02% level. This contrasts with the distribution of temperatures for these two classes, which do not show a significant difference. The immunoglobin-like (P.S.I) proteins and enzymes show an opposite behavior, where their distribution of temperatures is different at the 0.14% level of significance while there is no significant difference in the distributions of their pH values.

Table 4
Conditions generated by screen-design program for heme-containing proteins.

P.S.H, 294 K

*phosphate pH 6.8, 8.0% PEG 4000, iron	*phosphate pH 6.8, 25.0% PEG 1000, 0.3 *M* ammonium phosphate	Imidazole pH 6.6, 30.0% PEG 2000, iron	*phosphate pH 6.8, 10.0% PEG 6000, 0.05 *M* *citrate	*phosphate pH 6.4, 24.0% PEG 4000, iron	*phosphate pH 7.2, 20.0% PEG 2000, iron
*phosphate pH 6.4, 30.0% PEG 1000, 0.005 *M* KCN	*phosphate pH 7.2, 20.0% PEG 8000, 0.3 *M* ammonium phosphate	*phosphate pH 6.8, 20.0% PEG 8000, 0.020 *M* KCN	*phosphate pH 6.8, 18.0% PEG 6000, 0.3 *M* ammonium phosphate	*phosphate pH 8.0, 16.0% PEG 4000, iron	PIPES pH 6.4, 25.0% AS, 0.020 *M* KCN

Table 2 shows other interesting differences: membrane-associated proteins (P.M) are clearly outliers. Their distribution of reported temperatures is significantly different, often markedly so from that of almost every other major protein class. Heme/porphyrin-containing protein (P.S.H) crystallizations show a difference in their temperature distribution compared with that of enzymes. DNA-binding (P.DNA) and (small) ligand-binding (P.S.B) proteins are reported to crystallize with significant differences in the distribution of their pH values (at the 0.01% level).

The results shown in Table 2, together with other similar results for other comparisons (not shown), clearly provide objective support for both the idea that there are patterns of crystallization and the idea that a macromolecular classification scheme, such as that described here, captures some of those patterns. Unfortunately, however, these results do not provide much guidance 'at the bench' to someone trying to crystallize a new macromolecule.

We therefore developed the software described above with the goals of identifying probable crystallization conditions for a given macromolecular class and guiding actual experiments toward those conditions. A related goal was to facilitate the capture of crystallization data in machine-readable form so that failure data could be included in future versions of the program.

Below are presented sample results of calculations using this software. These are intended to illustrate the program and the underlying differences in the BMCD data. They do not necessarily represent generalized 'screens' for the macromolecular classes reported. For example, the number of trials shown here was reduced for economy of presentation; we would strongly suggest a larger number of trial conditions in a real experiment. (For example, a small experiment might invoke 72–96 conditions at room temperature and 48 at 277 K.) Indeed, the whole reason for developing a program is to enable individuals to tailor their crystallization screens to the specifics of their particular problem.

Table 3 shows the results obtained with the screen-design program for enzymes (P.S.E) using buffers, precipitating agents and their ranges, and salts and their ranges as reported in the BMCD. The actual choice of these values was facilitated by the program, which presents summary data for the selected class(es) in its display. The range for pH values was tailored to each buffer, spanning a range of $\pm$ 0.8 pH units on either side of the buffer's pK_a. Two temperature values were selected, 277 and 284 K (the binning range was $\pm$3 K). 1000 conditions were generated.

The table shows the 24 conditions with the highest probability score for 294 K. (24 was arbitrarily chosen because it represents a standard 'Linbro' crystallization tray.) The scores indicate that room temperature is preferred over cold (277 K) because the latter list had to be truncated at 12 experiments in order to terminate it at an estimated probability of success comparable with the 294 K results. This factor of two appears to propagate further; for example, a list of 48 conditions at 294 K terminates with approximately the same probability score as a list of 24 experiments at 277 K.

This preference illustrates an issue that arises whenever one bases a crystallization scheme on reported success rates – one cannot always readily separate human preferences from molecular behavior. For example, the dearth of reported crystallization successes at temperatures between 277 and 294 K almost certainly reflects human behavior. In this case, these two temperatures are readily available in most laboratories, while other temperatures are not. The statistical bias where significantly more enzymes have been reported to crystallize at room temperature than 277 K is more subtle. It is not clear whether this represents real molecular behavior or human reluctance to work in the cold.

Table 4 shows similar results for the heme-containing proteins (P.S.H) under conditions similar to those shown for the enzymes. Fewer conditions are shown because a series more appropriate to heme-containing proteins is shown below. These results show that the algorithm generates significantly different conditions for the two classes. For example, examination of the table shows that the reported pH range for heme-containing proteins is considerably narrower than that for enzymes.

The heme-containing protein results show an even stronger bias towards room temperature (with all the caveats discussed above) than do the enzymes. Indeed, the probability scores for heme-containing proteins at 277 K were so low that there were no experiments that correspond to those shown in Table 4; this is why the table only shows results for 294 K. However, longer lists (that delve lower in the probability scores) do pick up some 277 K experiments for heme-containing proteins, but the room-temperature conditions dominate. This differing distribution of reported results was indicated by the *t*-test results described above and this example shows how the program translates that difference into suggested screening conditions.

Table 5
Alternate conditions for heme-containing proteins.

PSHA, 294 K

*phosphate pH 7.2, 25.0% AS, 0.020 *M* KCN, iron	*Tris* pH 8.7, 25.0% AS, 0.10 *M* *acetate, ammonium phosphate	*phosphate pH 7.6, 25.0% PEG 1000, 0.005 *M* KCN, iron	*phosphate pH 8.0, 55.0% AS, 0.15 *M* *acetate, iron	*acetate pH 4.0, 25.0% PEG 1000, 0.010 *M* KCN, 0.3 *M* ammonium phosphate	*acetate pH 4.0, 10.0% PEG 6000, 0.020 *M* KCN, 0.3 *M* ammonium phosphate
Imidazole pH 6.6, 25.0% AS, 0.010 *M* KCN, ammonium phosphate	*cacodylate pH 6.3, 30.0% PEG 2000, 0.10 *M* *acetate, iron	*phosphate pH 6.4, 26.0% PEG 6000, 0.005 *M* KCN, iron	Imidazole pH 7.0, 4.0% PEG 4000, 0.40 *M* NaCl, 0.3 *M* ammonium phosphate	*phosphate pH 7.2, 30.0% PEG 6000, 0.020 *M* KCN, ammonium phosphate	*cacodylate pH 6.7, 35.0% AS, 0.10 *M* *acetate, iron
Imidazole pH 7.0, 45.0% AS, 0.05 *M* *acetate, iron	*phosphate pH 6.4, 10.0% PEG 6000, 0.35 *M* *acetate, ammonium phosphate	*phosphate pH 6.8, 26.0% PEG 6000, 0.020 *M* KCN, iron	*phosphate pH 6.4, 20.0% PEG 2000, 0.15 *M* *citrate, ammonium phosphate	*phosphate pH 6.8, 20.0% PEG 4000, 0.05 *M* KCN, ammonum phosphate	*phosphate pH 8.0, 28.0% PEG 8000, 0.015 *M* KCN, iron
*phosphate pH 8.0, 8.0% PEG 4000, 0.45 *M* *acetate, iron	*phosphate pH 7.2, 20.0% PEG 4000, 0.15 *M* *citrate, ammonium phosphate	PIPES pH 6.0, 25.0% PEG 1000, 0.10 *M* *acetate, iron	*phosphate pH 8.0, 12.0% PEG 8000, 0.20 *M* *acetate, iron	*phosphate pH 7.6, 10.0% PEG 6000, 0.30 *M* *citrate, iron	*phosphate pH 7.2, 20.0% PEG 4000, 0.20 *M* *acetate, iron

During the setup phase for the preceding example, it was obvious that one of two additives was used in virtually every reported successful crystallization of a heme-containing protein; they were ammonium phosphate and iron (II) citrate. They therefore dominated the 'additive' variable because of their statistical prevalence. However, the same displays also showed that other additives were usually present, *e.g.* salts such as KCN. This result shows the utility of the statistical summaries presented by the software. Accordingly, an additional run was performed using two additives: the first was a general salt, as used above, with the same trial values used for the enzymes, while the second additive was either ammonium phosphate or iron (II) citrate. These results are shown in Table 5 and clearly indicate how the program can be used to target the search on known properties of a given class of proteins. They are more appropriate for a heme-containing protein, although here too the list has been truncated for brevity. Significantly more experiments should be generated in a real crystallization attempt.

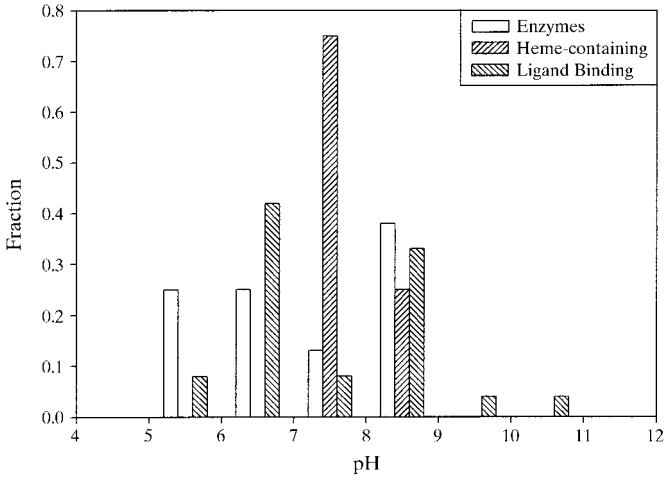

Figure 4
Distribution of pH values generated by the program for three classes of proteins, enzymes, heme- (porphyrin-) containing proteins and small ligand-binding proteins.

Table 6 shows typical results for the small ligand-binding proteins (P.S.B). Recall that the statistical (*t*-test) results suggested significant differences in the distribution of reported pH values for ligand-binding proteins and enzymes. This is borne out in the data, as can be seen in Fig. 4, which shows histograms of the top-scoring pH values suggested by the program. Although the pH distributions as represented by the histograms are similar, that of the ligand-binding proteins is clearly shifted towards higher pH values, with more high-scoring experiments at alkaline pH.

This shift is a clear illustration of what we are trying to achieve. While the ligand-binding proteins and the enzymes do not crystallize under radically different conditions, one can increase the efficiency of a search by shifting the sampling pattern of one's search; in this case, by shifting it to higher pH values. Subtler shifts in other parameters of the crystallization are also present.

These results also illustrate, sometimes by omission, the difficulties inherent in using a database grounded in the published literature. For example, we have not been able to include a consideration of the isoelectric point (pI) of the protein at this stage. It is well known that the solubility of most proteins is a minimum at their pI; it is therefore highly relevant to the design of crystallization conditions. However, publications describing preliminary crystallization results rarely (if ever) report the pI. Similarly, they virtually never report 'failure' data, *e.g.* conditions that gave rise to precipitates rather than crystals. Future versions of the program will

Table 6
Conditions generated for small ligand-binding proteins.

P.S.B, 294 K

HEPES pH 7.6, 60.0% As, 0.20 *M* NaCl	*citr* pH 5.7, 6.0% PEG 6000, 0.01000 *M* $CaCl_2$	*phosphate pH 7.2, 2.500% ethanol, 0.20 *M* *acetate	*Tris pH 9.1, 28.0% PEG 4000, 0.05 *M* NaCl	*citr* pH 5.3, 16.0% PEG 4000, 0.20 *M* NaCl	*acetate pH 5.6, 12.0% PEG 4000, 0.05 *M* NaCl
*acetate pH 5.6, 55.0% AS, 0.00250 *M* $CaCl_2$	MES pH 5.4, 14.0% PEG 6000, 0.15 *M* NaCl	*citr* pH 4.9, 14.0% PEG 6000, 0.01000 *M* $CaCl_2$	PIPES pH 7.2, 20.0% PEG 8000, 0.05 *M* NaCl	*Tris pH 8.3, 55.0% AS, 0.10 *M* *acetate	*acetate pH 5.6, 8.0% PEG 4000, 0.00500 *M* $CaCl_2$
MES pH 5.8, 12.0% PEG 8000, 0.15 *M* NaCl	PIPES pH 7.2, 20.0% PEG 8000, 0.01000 *M* $CaCl_2$	MES pH 5.8, 20.0% PEG 8000, 0.01000 *M* $CaCl_2$	*cacodylate pH 5.5, 4.0% PEG 4000, 0.01000 *M* $CaCl_2$	HEPES pH 7.2, 18.0% PEG 6000, 0.00500 *M* $CaCl_2$	MES pH 6.6, 4.0% PEG 8000, 0.35 *M* NaCl
*phosphate pH 7.2, 15.0% MPD, 0.05 *M* NaCl	*cacodylate pH 6.7, 24.0% PEG 8000, 0.05 *M* *acetate	*Tris pH 7.9., 8.0% PEG 4000, 0.25 *M* NaCl	HEPES pH 7.6, 25.0% MPD, 0.40 *M* NaCl	*acetate pH 4.8, 30.0% AS, 0.10 *M* *citrate	MES pH 5.4, 22.0% PEG 6000, 0.05 *M* *acetate

P.S.B, 277 K

Tris pH 8.7, 45.0% AS, 0.0001 *M* $MnCl_2$	*phosphate pH 6.8, 14.0% PEG 6000, 0.05 *M* KCl	*citr* pH 5.3, 25.0% MPD, 0.0001 *M* $MnCl_2$	*phosphate pH 8.0, 10.0% PEG 6000, 0.10 *M* NaCl	Imidazole pH 7.0, 5.000% ethanol, 0.05 *M* KCl	*phosphate pH 7.6, 45.0% AS, 0.0001 *M* $MnCl_2$
*phosphate pH 7.6, 10.000% ethanol, 0.00500 *M* $CaCl_2$	*cacodylate pH 5.5, 10.000% ethanol, 0.00500 *M* $CaCl_2$	*citr* pH 5.3, 15.0% MPD, 0.00500 *M* $CaCl_2$	*citr* pH 4.5, 5.000% ethanol, 0.00250 *M* $CaCl_2$	*acetate pH 5.2, 5.0% MPD, 0.05 *M* *acetate	*cacodylate pH 6.7, 7.500% ethanol, 0.05 *M* *acetate

include this information because we plan to gather our own data by recrystallizing a representative set of proteins where the relevant data will be gathered and recorded for this purpose. In the interim, the results and programs described here do use the data currently available in the BMCD to provide guidance during the initial screening for crystallization conditions.

Although there are limitations on the screen-design program described here, such as number of parameters it considers, we believe it provides useful information to crystallographers attempting to crystallize complex macromolecules. Preliminary results indicate that the program is capable of locating regions in a multi-dimensional space that are more likely to produce diffractible crystals. Moreover, the reasons for the program's conclusions are based on actual relative frequencies of success for different combinations of conditions as reported in the BMCD. The program's calculation of probability of success follows from a straightforward application of Bayes' theorem. As new experimental data is included, the probabilities will be calculated with increased confidence and the resulting screen designs should be even better.

This research is supported in part by funds from the W. M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University and the Pittsburgh Supercomputing Center, NIH National Center for Research Resources (NCRR) grant NIHRR10447 and National Library of Medicine Grant 2T15LM\DE07059.

References

Carter, C. W. Jr & Carter, C. W. (1979). *J. Biol. Chem.* **254**, 12219–12223.

Gilliland, G. C. (1987). In *Proceedings of the Second International Conference on Protein Crystal Growth, A FEBS Advanced Lecture Course.* Bischenberg, Strasbourg, France: North Holland.

Hennessy, D., Gopalakrishnan, V., Buchanan, B. G., Rosenberg, J. M. & Subramanian, D. (1994). *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, edited by R. Altman, C. Brutlag, P. Karp, R. Lathrop & D. Searls, pp. 179–187. Menlo Park, CA: AAAI Press

Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.

Samudzi, C. T., Fivash, M. J. & Rosenberg, J. M. (1992). *J. Cryst. Growth*, **123**, 47–58.