Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Machine-learning techniques for macromolecular crystallization data

**Vanathi Gopalakrishnan,[a,b] Gary Livingston,[a] Daniel Hennessy,[a] Bruce Buchanan[a] and John M. Rosenberg[c]***

[a]Intelligent Systems Laboratory, University of Pittsburgh, Pittsburgh, PA 15260, USA, [b]Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15260, 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 25 February 2004
Accepted 8 July 2004

Systematizing belief systems regarding macromolecular crystallization has two major advantages: automation and clarification. In this paper, methodologies are presented for systematizing and representing knowledge about the chemical and physical properties of additives used in crystallization experiments. A novel autonomous discovery program is introduced as a method to prune rule-based models produced from crystallization data augmented with such knowledge. Computational experiments indicate that such a system can retain and present informative rules pertaining to protein crystallization that warrant further confirmation *via* experimental techniques.

## 1. Introduction

Protein crystallization, like many aspects of structural biology, is becoming increasingly data-intensive. Data accumulate in many forms, including databases, the published literature and laboratory notebooks. Indeed, these data are so voluminous that automated methods for their interpretation have become mandatory. There are several informatics aspects of the crystallization problem that are especially challenging for reasons discussed below. One of these is to develop computational methods that effectively find useful associations present in the data, *i.e.* methods for data mining/knowledge discovery. Another is that much of the data are not collected in a form that is well suited for machine interpretation. For example, a list of reagents and concentrations in a crystallization cocktail can be very informative to a person reading a journal report, but to a computer 'ammonium sulfate', 'polyethylene glycol 4000' and 'polyethylene glycol 6000' are simply character strings that are devoid of chemical significance. Here, we report a method for importing that significance into two interrelated methods of machine learning. One of these methods was developed around the protein-crystallization problem. It is worth emphasizing that here we also report that progress in applying machine-learning methods requires simultaneous attention to both the computational and the descriptive aspects of the problem.

### 1.1. Background

Crystallization is the first step in structure determination by X-ray crystallography; it is typically also the rate-limiting step. Although the general physical-chemical theories that underlie crystallization are understood in principle, the detailed theory of the forces that promote and maintain macromolecular crystal growth is still preliminary. Most, if not all, macromolecular crystallization efforts are highly empirical, with each case being somewhat unique and idiosyncratic. Hence, it is necessary to search empirically for the optimum value of experimental conditions from a large space of about 25

parameters (*e.g.* temperature, pH *etc*.). This process is primarily trial and error, with each successive iteration leading to improvements in the size of the crystal and, more importantly, in the quality of its X-ray diffraction pattern. During the course of these experiments, the crystallographer accumulates substantial data on unsuccessful, partially successful and (hopefully) successful crystallization conditions.

Thus, there is a wealth of experimental data on successful and failed crystallization trials from which we can induce patterns or theories (correlations as well as causality) that capture relationships between experimental parameters, experimental protocols and protein characteristics. Such empirically derived theories can provide a rational approach to macromolecular crystallization and improve the probability of success of future crystallizations.

The Biological Macromolecule Crystallization Database (BMCD) is a database constructed by Gilliland (1988) that captures information about successful crystallization experiments. This database has been analysed several times to obtain an initial set of screening conditions for crystallizing a new macromolecule. Samudzi *et al.* (1992) performed a cluster analysis on version 1.0 of the BMCD and suggested a set of screening conditions specific to a major class of macromolecules. Gopalakrishnan *et al.* (1994) recreated these clusters using two kinds of methods: statistical analysis (similar to those of Samudzi *et al.*, 1992) and *COBWEB* (Fisher, 1987; a machine-learning and discovery program). The results from the clustering analysis were then used as input to the *RL* (Clearwater & Provost, 1990) inductive rule-learning program, resulting in verification and expansion of Samudzi's results (Hennessy *et al.*, 1994). Hennessy *et al.* (2000) augmented the BMCD with a hierarchical classification of the macromolecules contained therein, as well as data on the additives used with them and performed a statistical analysis that has led to a Bayesian technique for postulating the degree of success of a set of experimental conditions for a new macromolecule belonging to some known class.

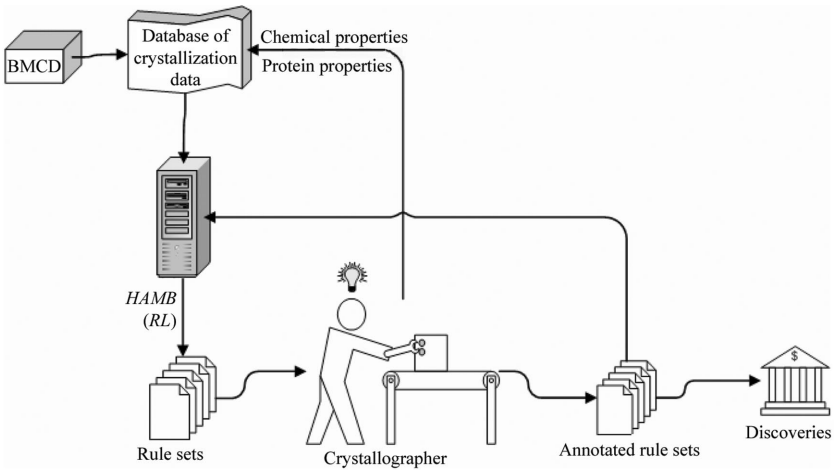


**Figure 1**
The iterative process of crystallization data evaluation reported here. The main loop includes evaluation and annotation of 'rule sets' produced by *HAMB* (see text), which are then used to generate heuristics used by *HAMB* to filter subsequent rule sets. With each iteration, the rule sets produced by *HAMB* reflect more closely the features of interest.

However, as noted by Jurisica *et al.* (2001), there are limitations in the BMCD. As the data are extracted from the literature, negative results are not reported in the database and many crystallization experiments are not reproducible owing to an incomplete method description, missing details or erroneous data. For instance, the complete list of chemical additives, an important factor in the ability to grow a crystal successfully, is not reported for many entries in the database. Furthermore, the crystallization conditions that were tried may in some cases have as much to do with the personal preferences of the investigators as they do with the chemical and physical requirements of the protein. It is generally impossible to discern between the personal preferences and 'real' chemistry from a purely retrospective look at the results. Nevertheless, it is clear that associations can be found within the BMCD data which can then be usefully interpreted. As more data become available in public databases, *e.g* the Protein Data Bank, or in electronic laboratory notebooks, these problems will tend to ameliorate, but will not disappear.

Data mining is meaningful only when there is a sufficient amount of useful data available for the purposes of statistical analysis and model building. What can be done in cases where data is limited and there is still an intrinsic need for better understanding of underlying phenomena within such a limited data set? One possible answer lies in feature construction or the building of an appropriate set of descriptors to augment the limited data set and facilitate statistical learning techniques to uncover underlying patterns within the data. The Biological Macromolecule Crystallization Database (BMCD) is a useful data set that provides a record of experimental conditions for successfully crystallized macromolecules. Yet, the number of descriptors and the depth of their descriptions are limited. As such, data mining of the BMCD yielded little success in terms of understanding the associations between the descriptors and the outcome of experiments, namely the kind and quality of crystal produced. In this paper, we present a hierarchical representation of additives and show the usefulness of augmenting a new of set of descriptors to the BMCD database in terms of the interestingness of the rules discovered by our novel program called *HAMB* (pronounced ham-bee; Livingston *et al.*, 2001*a*,*b*).

Simply adding more information in the form of new descriptors is insufficient to uncover useful underlying associations, as it leads to the production of a larger number of redundant or useless associations. Our encoding of the data in the form of hierarchies coupled with the incorporation of heuristic rules into the development of a novel prototype-discovery algorithm enabled the learning of associations from within the augmented data set. In this paper, we present some insights into these heuristic rules: how they worked to eliminate non-interesting associations and retain useful discoveries.

**Table 1**
Representation of a subset of features in one BMCD experiment entry.

| Attribute value and representation | Interpretation |
|---|---|
| ((macmol "Alcohol-dehydrogenase")<br>(maccon 5) (crmethod Bul-Dialysis)<br>(srctis liver) (srcgsp equus-caballus)<br>(pH 8.4) (temp 4.0) (buffer Tris)<br>(spacgp C222-1) (diflim 2.4)) | An entry in the BMCD that contains the successful crystallization conditions of alcohol dehydrogenase obtained from horse liver using the bulk-dialysis method. Favorable conditions were a macromolecular concentration of 5 mg ml$^{-1}$, a pH of 8.4, Tris buffer and a temperature of 277 K. The crystal that resulted from this experiment diffracted well with a diffraction limit of 2.4 Å and belonged to space group $C222_1$. |

## 2. Methods

Three basic methods form the foundation of this report. The first is an established machine-learning technique called *RL* (*Rule Learner*; Clearwater & Provost, 1990) that attempts to find associations in complex data that can be expressed as 'if–then' rules. Secondly, additional information is built into crystallization data to provide chemical significance to the list of additives in any crystallization experiment. This is achieved by hierarchical descriptors that are both categorical and numeric, as described below in §2.2. The third is a heuristic learning program termed *HAMB* that is built upon *RL*. *HAMB* initiates multiple runs of *RL*, directing it, filtering its output, grouping results and presenting them to the user in useful ways, where 'usefulness' is one of the things determined by the heuristics (Fig. 1).

### 2.1. Inducing rules from data with *RL*

*RL* (Clearwater & Provost, 1990) is an inductive learning program that learns by incrementally generating rules and testing them against the available data. It was first used to learn rules for predicting mass spectra of complex organic molecules (Feigenbaum & Buchanan, 1993). A rule consists of a set of conditions and a predicted outcome (*i.e.* if ⟨conditions⟩ then ⟨outcome⟩), so testing the rules consists of finding all of the data that meet the conditions and computing the percentage of those data that match the predicted outcome. New rules are generated by specializing existing rules, *i.e.* by adding new conditions, with the goal of filtering out data that did not match the predicted outcome while retaining the data that did match the outcome. The result is a process that incrementally focuses the rules on patterns that exist in the data.

One feature of *RL* that makes it a flexible learner is its ability to use background knowledge to constrain the search for rules. Background knowledge includes such information as constraints on numeric valued attributes, such as range of value and step size, desirable properties of rules being learned, as well as available ISA hierarchies of domain attributes. The ISA hierarchy represents the 'kind-of' relationships between attributes, such as 'lysine' ISA (is a kind of) 'polar amino acid'. This latter property of being able to constrain the search space of hypotheses of rules using ISA hierarchies is very useful for our purposes of representing and dealing with macromolecular class hierarchies and hierarchy of additives. The hierarchy of space groups associated with crystal Bravais lattice structures constitutes an ISA hierarchy: for example, $P6_222$ ISA $P6_2$ ISA $P6$ (space group $P6_222$ is subfamily of $P6_2$, which in turn is a subfamily of space group $P6$, a hexagonal class system). New meaningful attributes were added to enhance the readability of the rules that were output by *RL*. For instance, an individual space group was converted to represent four different attributes; namely, crystal system, point group, centering and polymorphic. The crystal system values were based on the Bravais lattice.

*RL* has statistics associated with each rule produced from the training data. Below is an example of a sample rule discovered by *RL* indicating how likely, given the data, is it that a crystal with habit 'plates' would have a resolution limit of diffraction under 3.5 Å.

$$\text{CRHABIT PLATES} \rightarrow \text{DIFLIM-UNDER-3.5}$$
$$p = 45, n = 5, tp = 520, tn = 109.$$

Here, $p$ and $n$ are the number of positive and negative examples in which this rule applies and $tp$ and $tn$ are the total number of true positive and true negative examples available. Each rule also has with it numbers indicating sensitivity, specificity and positive predictive value.

### 2.2. Representation of additives and chemical relationships

In order to use *RL* to analyze the BMCD data, we had to convert it into a usable form (Table 1). This reorganization of the data is described in Hennessy *et al.* (1994). In this section, we describe further augmentations to the BMCD to enhance *RL*'s analysis of the data.

Considering ammonium sulfate reveals the problems inherent in describing crystallization additives to a computer. It is typically used in 'high' concentrations (over 1 *M*) as a precipitating agent. However, it is obviously a salt that, for example, would significantly alter the ionic strength, especially at high concentrations. Indeed, it is the salt of a strong acid and a weak base; hence, it also acts as a buffer at alkaline pH. All these factors must be considered if one wants to objectively assess relationships in crystallization data.

We therefore developed two independent hierarchical schemes to describe each additive. The first relates to the 'commonly perceived role' of precipitating agent, salt, buffer *etc.*, because this is how crystallization experiments are usually presented and discussed in the literature. The second provides a framework for the description of the physical and chemical properties of each additive. Many, like ammonium sulfate, dissociate into species in solution; in this example, it dissoci-

ates into $NH_4^+$ and $SO_4^{2-}$. The basic descriptor for each additive is simply a list of the corresponding species found in solution. This is followed by descriptors of each species, including charge and p$K_a$, as indicated below. This facilitates a global description of the mother liquor. For example, if the buffer is ammonium phosphate, ammonium ions would also be contributed from this source; this approach facilitates a simple calculation of the combined concentration. It also facilitates the calculation of overall properties, such as ionic strength.

We group the information contained in our augmented version of the BMCD into five groups. Note that while some of this information was in the BMCD, most of it was not.

(i) Additive properties: information pertaining to additives, such as their concentrations and perceived roles in the cases.

(ii) Species properties: information about the chemical species that the additives break down into, such as their chemical classifications and concentrations.

(iii) Global properties: information about the crystal-growth experiment as a whole, such as the crystallization method, pH of the solution and temperature.

(iv) Macromolecular properties: properties of the macromolecule being crystallized, such as its molecular weight or classification.

(v) Crystal properties: properties of the resulting crystal, such as its diffraction limit and form.

**2.2.1. Description of additive properties**. We represent the chemical properties of additives by their concentrations and perceived role. An additive's perceived role is the typical role in which an additive is used; for example, polyethylene glycol is typically used as a precipitating agent. The perceived roles form a hierarchy and are presented in Fig. 2. The BMCD provides information on the additives present in a solution along with their concentration.

Our machine-learning program requires that the data be given in row–column format, with each row representing an experiment and each column representing an attribute (*i.e.* property or characteristic) of that experiment. Therefore, each experiment has the same number of attributes. However, in the BMCD there are a varying number of additives present. We represent each additive used in a reasonable number (five) of experiments in the original BMCD as a separate attribute whose value refers to its concentration. For our initial research purposes, we do not consider rarely used additives, as the value of concentration for such attributes would be zero for most cases and is, therefore, unlikely to be statistically significant. In the future, we would like to implement different strategies for attribute selection in order to overcome the limitations that could arise from this simple assumption. Now that crystallization data is being entered into the PDB, it will enable us to employ more sophisticated feature selection.

Precipitating agent
Buffer
Salt
Detergent
Substrate analogue
Heavy-metal reagent
Not applicable, not available, unknown
Preservatives:
- Miscellaneous organic
- Chelator
- Reducing agent
- Anti-microbial
- Anti-protease

**Figure 2**
Hierarchy showing the commonly perceived roles of additives.

Within the BMCD data set, some concentrations were specified as molarities, while others were given as percentages. We converted concentrations given in percentages to molarities when possible. If it was not possible to make this conversion, we simply represent the additive as being present. Because 47 additives were used in more than five experiments, we added 47 attributes to the BMCD, with possible values of either present, meaning the corresponding additive was present but the concentration could not be converted to molarity, or numbers representing the concentrations (or absence) of their respective additives.

Equations (1) to (4) depict our formulas for converting percentage values to molarity. We use linear approximations to represent the relationships. For ammonium sulfate, percentage saturation is converted to molarity as

$$\text{Molarity} = \frac{\text{Percentage saturation} \times (4.05\ M)}{100}. \quad (1)$$

For solids specified in percent weight per volume (*w*/*v*), we calculate molarity as

$$\text{Molarity} = \frac{\text{Percentage concentration}}{100}[\text{g (100 ml)}^{-1}] \times (1000\ \text{ml l}^{-1}) \times \left[\frac{1}{\text{mol wt (g } M^{-1})}\right] \quad (2)$$

For liquids, we convert percentage volume to molarity. For liquids with specific gravities,

$$\begin{aligned}\text{Molarity} &= \frac{\text{Percentage concentration}}{100} \times (1000\ \text{ml l}^{-1}) \times \frac{\text{specific gravity (g ml}^{-1})}{\text{mol wt (g } M^{-1})} \\ &= \frac{\text{Percentage concentration} \times 10 \times \text{specific gravity}}{\text{mol wt}}.\end{aligned} \quad (3)$$

For liquids with relative densities, molarity is calculated as

$$\begin{aligned}\text{Molarity} &= \frac{\text{Percentage concentration}}{100} \times (1000\ \text{ml l}^{-1}) \times \frac{\text{relative density (g ml}^{-1})}{\text{mol wt (g } M^{-1})} \\ &= \frac{\text{Percentage concentration}}{100} \times (1000\ \text{ml l}^{-1}) \times \frac{\text{density (g ml}^{-1})}{\text{mol wt (g } M^{-1})} \\ &= \frac{\text{Percentage concentration} \times 10 \times \text{density}}{\text{mol wt}}.\end{aligned} \quad (4)$$

- Inorganic
  - Alkaline metal, e.g. $Na^+$
  - Carbonate
  - Divalent metal (2nd row), e.g. $Mg^{2+}$
- Halide, e.g. $Cl^-$
  - Nitrate
  - Other
  - Phosphate
  - Sulfate
- Organic
  - Aldehyde
  - Aliphatic
  - Amide
  - Amine
  - Aromatic
  - Arsene
  - Carboxylic acid
  - Ether
  - Organic chelator
  - Sulfhydryl
  - Sulfonate
  - Organic hydroxyl
    - Polyalcohol
    - Simple, single hydroxyl, e.g. ethanol
- Detergent
  - Ionic detergent
  - Nonionic detergent
- Metabolites: substrates, cofactors, analogues, etc; things that might bind to a protein and/or DNA
  - Amino acids and their analogues
  - Dinucleotide
  - Glyceride
  - Nucleic acid bases
  - Nucleoside
  - Nucleotide
  - Other - something else that may (or may not) bind
  - Sugar-phosphate
  - Sugars and simple carbohydrates
- `Heavy atom' reagents usually used for isomorphous derivatives

(*a*)

**By $pK_a$**
- Strong acid – has a $pK_a$ less than or equal to 3.5
- Acid – has a $pK_a$ between 3.5 and 6
- Neutral – has a $pK_a$ between 6 and 8
- Base – has a $pK_a$ between 8 and 10.5
- Strong base – has a $pK_a$ greater than 10.5

**Number of titrateable groups**
- Titrateable groups $\geq 5$
- Titrateable groups = 4
- Titrateable groups = 3
- Titrateable groups = 2
- Titrateable groups = 1
- Titrateable groups = 0

**By polymerization state**
- Two monomers (i.e. a dimer)
- A few monomers – (i.e. a trimer or tetramer)
- Polymer - many monomers

**By net charge**
- Highly negative – having a net charge $\leq -5$
- Charge = −4
- Charge = −3
- Charge = −2
- Charge = −1
- Charge = 0
- Charge = 1
- Charge = 2
- Charge = 3
- Charge = 4
- Highly positive – having a net charge +5
- Dipole (i.e. a zwitterion)
- Somewhat-mixed charge – a mixed distribution of positive, negative and neutral charges

(*b*)

**Figure 3**
(*a*) Hierarchy showing the species present in additives. (*b*) Classifications of the species based upon select chemical properties.

Similarly, we added 11 attributes to represent the perceived roles that were used in a reasonable number (five) of experiments in the BMCD. Each attribute represents the total concentration of additives filling a corresponding perceived role. We convert all concentrations to molarity to facilitate the computation of the concentrations, since concentrations expressed by molarity may simply be summed to compute the total concentration.

**2.2.2. Description of species properties**. Figs. 3(*a*) and 3(*b*) present our hierarchical classification of the species. To represent this classification in our augmented database, we flattened the hierarchy, using one attribute for each group in our classification. The value of the attribute is either (i) the total concentration of species in the crystallizing solution that belong to the corresponding group, if all concentrations are in molarity, or (ii) 'present', if precise numerical information is not present for all cases in which the attribute appears. For all groups having no species present in the solution, the value is 0. For example, suppose there are only ammonium, sodium and spermine ions present at concentrations of 1, 3 and 1 *M*, respectively. Then, the concentrations of the o.amine and i.alk groups would be 2 and 3 *M*, respectively. For all other groups, the concentrations would be 0. 23 of the groups were used in more than five experiments in the BMCD; therefore, there are 23 attributes in our augmented BMCD, representing the concentrations of the species belonging to the groups. The coverage value of five experiments was chosen as a reasonable starting point since RL retains only those rules that cover a

**Table 2**
Sample input file data specified to *HAMB*.

Each example (experiment in BMCD) is specified to the program as a row consisting of values for each attribute. Since the number of attributes is large, we present the inverted table to allow several important attributes to be shown.

| Attributes | Example 1 | Example 2 | Example 3 | Example 4 |
|---|---|---|---|---|
| id | 2360 | 2358 | 2356 | 2353 |
| sodium-acetate-concentration | not-present | not-present | >0.05<= 0.1-m | not-present |
| sodium-chloride-concentration | not-present | not-present | not-present | >0.05<= 0.1-m |
| role-salt-concentration | not-present | not-present | >0.05<= 0.1-m | >0.05< 0.1-m |
| phosphate-concentration | not-present | >1.3-m | not-present | not-present |
| sodium-concentration | not-present | >0.3529-m | >0.04<= 0.1-m | >0.04<= 0.1-m |
| diffraction-limit | ? | <= 2 | ? | >2.4& = 2.8 |
| buffering-capacity | ? | >0.036 | >0.004&<= 0 | >0.004&<= 0.036 |
| Ionic strength | <= 0.025 | >5.975 | >0.025&<= 0.157 | >0.157&<= 2.207 |
| Macromolecule-class | p.s.l.pep | p.s.e | misc | p.s.i |

reasonable number of training cases and are therefore more likely to be statistically significant when learning a general concept.

The ionic strength is calculated whenever all concentrations of the species in the solution were able to be represented in molarities. For each species $i$ in the solution, we first calculate the root-mean-square concentration for that species as

$$\text{RootMeanSqConc}_i = \begin{cases} \text{minimum conc}_i & \text{if max = min conc}_i \\ 0.5(\text{minimum conc}_i^2 + \text{maximum conc}_i^2) & \text{if conc}_i \text{ is not constant.} \end{cases} \quad (5)$$

The ionic strength is then calculated using the formula

$$\text{ionic strength} = \frac{\sum(\text{RootMeanSqConc}_i \times \text{Charge}_i^2)}{2}. \quad (6)$$

We also grouped the species by many of their chemical properties such as p$K_a$ values, titratable groups, polymerization and net charge (see Fig. 3*b*). As with our classification groups, for each of these groupings the value of the corresponding attribute was the total concentration of all species in the crystallization experiment that belong to that group, as long as all concentrations are expressed in molarity; otherwise, the value for that attribute is present. When all concentrations for species in the solution are represented using molarities and the p$K_a$ values for all the species are known, we calculate buffering capacity as the slope of the titration curve from the Henderson–Hasselbach equation,

$$\text{buffering capacity} = \log_e 10 \times \left\{ \sum_{s \in \text{species}} \sum_{\text{p}K_a \in \text{p}K_a s} \frac{\text{MeanConcentrations} \times (\text{pH} - \text{p}K_a)^{10}}{[1 + (\text{pH} - \text{p}K_a)^{10}]^2} \right\}. \quad (7)$$

**2.2.3. Description of the global properties of crystallization solutions**. These properties pertain to the crystallization experiment as a whole. The entry number (which we ignored), crystallization method description, pH and temperature were provided in the original database. We added information about the crystallization method scale, crystallization method type and, now that we have the necessary chemical information, the buffering capacity and ionic strength of the crystallization solution.

**2.2.4. Description of the macromolecular class**. The original BMCD contained the macromolecule ID, macromolecule weight and macromolecule concentration. We added an attribute representing a hierarchical classification of the macromolecules, which is detailed in Hennessy *et al.* (2000).

**2.2.5. Description of crystal properties**. These properties consist of the diffraction limit and the crystal form. For the latter, we used a hierarchical description of Bravais lattice space groups along with the authors' description of the crystal habit, which is usually obtained by visual inspection.

To summarize, the attributes in our augmented data set include the following.

(i) Macromolecular properties: macromolecule name, macromolecule class name and molecular weight.

(ii) Experimental conditions: pH, temperature, crystallization method, macromolecular concentration and concentrations of chemical additives in the growth medium.

(iii) Characteristics of the grown crystal (if any): descriptors of the crystal's shape (*e.g.* crystal form and space-group description) and its diffraction limit, which measures how well the crystal diffracts X-rays.

### 2.3. Machine learning *via HAMB*

The output from *RL* (applied to the augmented data) was voluminous, in our case consisting of several hundred individual rules, each of them requiring a judgement as to their significance. One can tweak the parameters for *RL* in order to reduce the volume of the output, but that is time-consuming and tedious. This led eventually to the conception and development of a novel prototype autonomous discovery program called *HAMB* (Livingston *et al.*, 2001*a*,*b*) that reasons and estimates the interestingness of the patterns produced from the application of rule-based machine learning to the augmented crystallization database. HAMB encodes domain-specific knowledge and general heuristics that enable

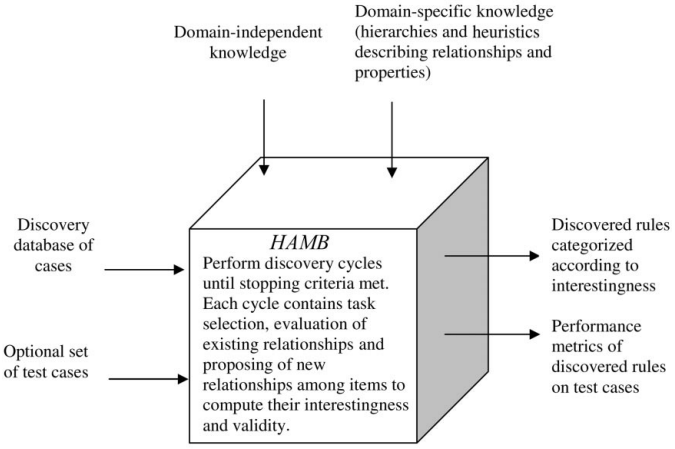


**Figure 4**
*HAMB* overview.

it to perform autonomous discovery from complex real-world data.

In Livingston *et al.* (2001*a*,*b*), we describe the *HAMB* program, which can decide for itself which discovery tasks to perform and when to perform them. *HAMB* utilizes user preferences and a small set of known relationships between the attributes in the data to automatically set up the *RL* program and to post-process the induced rules. *HAMB* consists of an agenda-and-justification-based framework for selecting the next task to perform. Tasks refer to computational encoding of operations on items that refer to instances of the search space of possible discoveries. Tasks are performed using heuristics that create new items for further exploration and that place new tasks on the agenda. This framework has several desirable properties: (i) it facilitates the encoding of general discovery strategies using various types of background knowledge, (ii) it reasons about the appropriateness of the tasks being considered and (iii) it tailors its behavior toward a user's interests by prioritizing tasks according to an estimate of interestingness specified by the user.

For example, a task in *HAMB* would be to 'examine the relationship between ionic strength and predictivity of good crystal'. Performing this task would involve the production of new sub-tasks such as 'induce rule-set' that will cause *HAMB* to set up the induction task by (i) selecting a training set of examples, (ii) selecting the feature set of attributes from which the rules will be induced, (iii) selecting the parameters with which to run *RL* and finally (iv) running *RL* to induce the rules. *HAMB* then loads the induced rules and post-processes them. The $p$ value of a rule's positive predictive value is calculated using Fisher's exact test using the $2 \times 2$ table generated by the number of true and false positives as well as negative test case predictions (based on a validation set of examples that were unseen during learning) for each rule, omitting those cases where there is no prediction. Fisher's Exact Test (FET) is a statistical test used to determine if there are nonrandom associations between two discrete variables. *HAMB* uses the $p$ values returned by FET in order to rank the rules, so that it may be able to decide in what order to present tasks to *RL*. Application of *HAMB* to crystallization data shows its power in identifying patterns that are both interesting and novel (Livingston *et al.*, 2001*a*,*b*).

As shown in Fig. 4 and Table 2 and in the supplementary material[1], *HAMB*'s input consists of the files containing the set of cases that it will use to make its discoveries (the discovery database), an optional testing set of examples (the testing database), a domain theory file containing the domain-dependent knowledge and files containing definitions of domain-independent relationships, properties, task types and heuristics. HAMB reports as discoveries those items with interesting relationships or properties, *i.e.* if its value exceeds a threshold provided for each relationship or property. If the testing database is not given to *HAMB*, it creates its own set of test instances comprising of a random one-third of the discovery database's cases (examples). One of *HAMB*'s methods for post-processing rules is to group them into rule families (see supplementary material). These are groups of rules where changing the value of one attribute on a rule's left-hand side results in a consistent change in the value being predicted (that is, the right-hand side of the rule). Apart from making it easier for the user, the consistency of rules within a family increases confidence in the rules themselves.

A major advantage of this framework is that it provides a clean separation of the discovery program from the knowledge it uses. We provide further modularity in *HAMB* by using domain-independent heuristics (and properties and relationships) which refer to domain- and problem-specific information that is either given in a domain theory file or discovered by *HAMB*. While *HAMB* and its heuristics are general, they access domain- and problem-specific information. Therefore, *HAMB* is able to perform discovery using domain-specific knowledge, allowing *HAMB* to tailor its behavior to the discovery problem and to evaluate the cases given to it to make discoveries from and the resulting discoveries in a domain-specific context. Thus, *HAMB* is able to examine our augmented BMCD database using a wide variety of knowledge specific to macromolecule crystallization. In contrast, most other knowledge-discovery, data-mining and machine-learning programs are only capable of using one or two types of background knowledge. Results of experiments with *HAMB*, reported in §3, demonstrate that it uses background knowledge effectively to evaluate its discoveries and avoid reporting a large number of uninteresting rules.

The types of domain-specific knowledge that *HAMB* uses are the following.

(i) Simple semantic information, such as potential target attributes and the value used to denote missing values. Target attributes refer to those class variables that are to be predicted, such as DIFLIM_UNDER_3, which refers to the class of crystals produced with a resolution limit of diffraction of less than 3.0 Å. There are several missing values for attri-

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: AV5008). Details for accessing these data are given at the back of the journal.

butes within the BMCD and a computer program such as *HAMB* needs to know how such missing values are represented. By allocating a character such as '?' to indicate a missing entry and letting *HAMB* know this to be the case, it is possible to handle this incomplete information appropriately. In most cases, *HAMB* will ignore this field during calculations. In cases where a concentration is specified, but the actual chemical name is missing, *HAMB* will treat the concentration as a missing value as well since these are related fields in the database.

(ii) Declarations of item groups, which is used to convey the user's *a priori* interests. *HAMB* provides two mechanisms for factoring a user's preferences into an item's estimated interestingness: (*a*) *via* user-defined item groups, sets of items to which the user assigns a utility, and (*b*) the weights used in *HAMB*'s function for estimating interestingness (see below). To define an item group, the user provides a name for the group, a utility (a relative number) and a predicate for determining the group's membership. Here, the user is more interested in attributes that describe the outcomes of experiments (that is, observables such as crystal type or crystal shape) than attributes that describe the characteristics of macromolecules (that is, givens such as the pI of protein) and is also relatively more interested in such given attributes as opposed to experimental controllable attributes such as the presence of additives. To model these interests, three item groups could be defined with utilities of 100, 50 and 25, respectively. This will bias *HAMB*'s discoveries to retain most relationships that contain observables as the predicted variable.

(iii) Interestingness weights: the weights *HAMB* uses to estimate the interestingness of an item. *HAMB* estimates the user's interest in an item using a hierarchical weighted sum of selected properties. The user provides these weights in the domain theory and they represent a crude model of the user's preferences. The estimation of the interestingness weights is made using abstraction and normalization as follows: (*a*) values of an item's properties are calculated, (*b*) the values are combined using weighted sums, (*c*) the values of the abstractions are normalized to fit between 0 and 100 and (*d*) interestingness weight estimates are computed using a weighted sum of these normalized top-level abstractions. For example, when estimating the interestingness of a rule, the properties positive predictive value, negative predictive value, ruleset-usage ratio, specificity and sensitivity are abstracted into a higher-level property called empirical support. Other higher-level properties such as 'novelty', 'semantic simplicity', 'syntactic simplicity' and 'utility' (see below) are derived from ten properties used to describe attributes. A more detailed example of the weights and abstractions can be found in Livingston *et al.* (2001*a*,*b*).

(iv) Relationships between the attributes and their values that are definitional or strongly believed to hold. *HAMB* uses these relationships to identify findings that are rediscoveries, avoiding presenting these findings as discoveries and to reduce the amount of redundancy in its reported discoveries. The user may specify a variety of known relationships among attributes and their values (see supplementary material) as follows below.

(*a*) Definitionally related. This one-to-many relationship is used when the first attribute is derived from other attributes. For example, ionic strength would be definitionally related to concentration and charge. In the supplementary material, examples of this relationship are shown between molecular weight of macromolecule and its class.

(*b*) Translational equivalence. This one-to-one relationship is a specialization of definitionally related and is used when two attributes measure the same feature, such as the two attributes *C* and *F*, which measure the temperature of the same object, but represent the Celsius and Fahrenheit temperature scales.

(*c*) Semantic equivalence. This relationship is used when two attributes should have identical values. The attribute names or descriptors are basically aliases for describing the same value. For example, PEG concentration and precipitant concentration would be equivalent if PEG was the precipitating agent.

(*d*) Abstraction. This relationship implies that an attribute *A* refers to the description of a more general class than another attribute *B*. For example, *A* refers to 'organic chemical compound' and *B* refers to a subfamily of organic compounds such as 'aldehyde'. For more examples, see Fig. 3(*a*). Several examples of abstraction relationships within macromolecular class, crystallization methods and space groups are depicted in the supplementary material.

(*e*) Discretization. Attributes that represent real values can be represented as discrete valued ranges: for example, PEG concentration can be represented as ranging from 0–10, 11–20, 21–30%(*w*/*v*); that is, in discrete ranges of 10 up to 100%.

(*f*) Known related. This relationship is used when one attribute is known to be related to a second attribute in a manner not expressible using the other relationships, such as when the lore of the domain is that the first attribute is somehow related to the second attribute but the exact relationship is unknown. For example, it has been found empirically that presence of small polyamines such as spermine in the crystallization solution is associated with high-quality nucleic acid crystals. This is clearly related to charge neutralization. *HAMB* found a similar association with $Mg^{2+}$ and thus can deal with situations where multiple chemical means can be used to achieve the same goal.

In the supplementary material, several examples of known associations between ionic strength and chemical species, buffering capacity and species are depicted.

## 3. Results and discussion

### 3.1. Use of *RL* on the BMCD data

We initially processed 1025 entries from an earlier version of the BMCD and ran *RL* on several subsets of the data divided based on three resolution limits of diffraction, namely 2.5, 3 and 3.5 Å (Hennessy *et al.*, 1994). For each separate run

of *RL*, we treated experiments within the BMCD reporting resolution limit above each of these values to be negative examples of good crystals (the split of positive to negative examples was approximately 75 to 25%, but varied depending on our subsetting criteria). We further focused on obtaining rules relevant to protein crystallizations alone. Our initial attempts at applying machine-learning techniques to the BMCD were partially successful in that we were able to induce rules that were indicative of the types of experimental conditions that were reported as resulting in high-quality crystals. We categorized the rules obtained from *RL* along two dimensions: (i) the kinds of associations between given, controllable and observable parameters of crystallization experiments and (ii) the rule content (*e.g.* known facts in crystallography, new relationships discovered). Based on this categorization, we were able to confirm several known relationships between given, controllable and observable parameters. *RL* also found several interesting associations, including those between buffers and their most effective pH value ranges from the BMCD, space groups and warmer temperatures for successful crystals and certain orthorhombic space groups as indicators of good crystal quality.

The results of our initial work in applying *RL* to the BMCD is described in Hennessy *et al.* (1994). The relationships reported from this analysis clearly indicated the need for the following: enriching the data representations, extending machine-learning techniques to work with these representation and incorporating additional forms of domain knowledge to guide induction. Furthermore, it was fairly clear that *RL* generated 'better' rules than those generated by statistical clustering analysis (Gopalakrishnan *et al.*, 1994; Samudzi *et al.*, 1992), in that the rules provide greater insight into the details of the discovered relationships.

### 3.2. Initial use of *HAMB* on augmented BMCD data yielded the expected information

We obtained 2225 examples from an updated BMCD database (Gilliland, 1988). These data were supplemented, as described above, with additional chemical information. The number of features of crystallization data in this new database grew to 170 descriptors. The additional information contains many known dependencies which a discovery program could find and report as discoveries, when actually they are not. Thus, this additional information is a double-edged sword: while it adds information to the database which may be useful to a discovery program, it also increases the redundancy and the number of non-novel patterns in the database, which can make it difficult to inspect the discoveries to identify the interesting discoveries, as well as lead to overfitting (Mitchell, 1997).

**Table 3**
Categorization of the interestingness of 575 discoveries made by *HAMB* from the augmented BMCD data.

The redundant rules removed during the semi-manual filtering (approximately 144) are counted as Category 0 discoveries. Removing the 144 rules from the calculations results in only 22% (96/431) Category 0 discoveries.

| Category | Description | Number | Percentage |
|---|---|---|---|
| IV | Individually, Category IV discoveries could be the basis of a publication in the crystallography literature, being both novel and extremely significant to crystallography | 0/575 | 0 |
| III | In groups of about a dozen, Category III discoveries could form the core of research papers in the crystallography literature if substantiated by further experiments | 92/575 | 16 |
| II | Category II discoveries are about as significant as Category III, but are not novel | 192/575 | 33 |
| I | Category I discoveries are not as interesting as Category II or III, but still are of some interest | 51/575 | 9 |
| 0 | Category 0 discoveries are any discoveries that are not Category I, II, III or IV | 240/575 | 42 |

### 3.3. Machine learning *via HAMB*

In addition to the augmentations to the BMCD, we provided *HAMB* with a wealth of chemical information about the attributes, in the form of a table of 1549 known relationships between the attributes. Excerpts of the information we provided to *HAMB* are provided in the supplementary material.

*HAMB* took approximately 24 h and 31 000 discovery cycles on a Compaq NT workstation running at 1.3 GHz with 128 MB RAM to process the augmented BMCD data. During that time, HAMB reported 575 discoveries. We categorized these discoveries by their significance and novelty. Our categories are shown in Table 3, along with the numbers (and percentages) of discoveries in each category. A few of the more interesting discoveries likely to be novel and significant are presented in Table 4. These discoveries may be helpful to crystallographers, but because the data are noisy and are biased by human preferences, further investigation is needed to confirm their validity. The first three rules in Table 4 suggest that different crystallization methods should be used for specific types of macromolecules. The last three rules in Table 4 suggest that different ionic strengths may be required when crystallizing enzymes, 'heme'-containing proteins and small proteins.

Table 5 presents a sample of *HAMB*'s (re)discoveries that represent associations that are known in the 'lore' that a second-year graduate student of structural biology might be expected to know. Some of these rules also reflect clear chemical reasoning. Divalent cations are needed to stabilize nucleic acid crystallization. Some of the other rules may have more intricate interpretations. For example, magnesium chloride is only rarely used in protein crystallization, possibly because of its well known tendency to form inorganic crystals. There may also be examples of 'cultural patterns', such as the nucleic acid community's use of cacodylate buffers. However, it remains to be seen whether cacodylate may have some kind of stabilizing role in nucleic acid crystallizations.

**Table 4**
Novel discoveries from *HAMB*.

The statistics reported for each rule are calculated from a validation set not used to learn the rule. The *p* value of a rule's positive predictive value is computed using the Fisher's exact test (Sokal & Rohlf, 1969). P.RNA.E macromolecules are proteins that bind to RNA and catalyze a chemical reaction that modifies it, P.S.H ('heme'-containing) macromolecules are soluble proteins containing an iron-porphyrin prosthetic group (*e.g.* hemoglobin and cytochrome), P.S.L macromolecules are small proteins and peptides and P.S.L.O macromolecules are heterogeneous subgroups of P.S.L.

| | True positives | False positives | Sensitivity | Positive predictive value | *p* value |
|---|---|---|---|---|---|
| Macromolecule class is P.RNA.E → crystallization method is batch | 141 | 88 | 0.39 | 0.62 | <0.001 |
| Macromolecule class is P.S.H → crystallization method is temperature-crystallization | 22 | 30 | 0.73 | 0.42 | <0.001 |
| Macromolecule class is P.S.L.O → concentration by evaporation | 67 | 14 | 0.65 | 0.83 | <0.001 |
| Macromolecule class is enzyme → ionic strength is greater than 2.21 and less than or equal to 5.98 | 151 | 638 | 0.53 | 0.21 | <0.001 |
| Macromolecule class is P.S.H → ionic strength is greater than 5.98 | 90 | 139 | 0.28 | 0.39 | <0.001 |
| Macromolecule class is P.S.L → ionic strength is less than or equal to 2.21 | 114 | 137 | 0.32 | 0.45 | <0.001 |

**Table 5**
A sample of discoveries from *HAMB* in which the associations are well known.

The association between nucleic acid crystallization and magnesium chloride used to stabilize the highly negative charge is rediscovered by *HAMB*. Divalent cations (including magnesium chloride) effectively stabilize nucleic acids because of their chemical properties. Many salts of divalent cations are insoluble, making them undesirable for protein crystallizations. Such patterns are depicted in the rules discovered by *HAMB* and increase our confidence in *HAMB*'s ability to detect patterns. Additionally, although the association between these properties is not novel, the numeric limits that *HAMB* also provides, such as those on the concentrations, can provide significant new information.

| | True positives | False positives | Sensitivity | Positive predictive value |
|---|---|---|---|---|
| The macromolecule is a nucleic acid → magnesium chloride is present with a concentration less than or equal to 0.04 *M* | 60 | 54 | 0.46 | 0.53 |
| The macromolecule is a nucleic acid → inorganic divalent species are present with a total concentration between 0.0025 and 0.04 *M* | 81 | 33 | 0.35 | 0.71 |
| The macromolecule is a nucleic acid → species with net charges ≥2 are present with a total concentration between 0.005 and 0.03 *M* | 64 | 50 | 0.43 | 0.56 |
| The macromolecule is a protein → inorganic divalent species are not present | 1660 | 262 | 0.92 | 0.86 |
| The macromolecule is a protein → magnesium chloride is not present | 1862 | 60 | 0.90 | 0.97 |
| The macromolecule is a protein → the species cacodylate is not present | 1857 | 65 | 0.90 | 0.97 |
| The macromolecule is a protein → species with net charge ≥2 are not present | 1274 | 648 | 0.92 | 0.66 |
| The macromolecule is a protein → spermine is not present | 1921 | 1 | 0.87 | 1.0 |
| The macromolecule is a protein → spermine tetrahydrochloride is not present | 1918 | 4 | 0.87 | 1.0 |

The ability of *HAMB* to discover well known associations from the augmented BMCD serves to both confirm the existence of useful patterns within this augmented database, as well as increase our confidence in the validity of patterns learned by *HAMB*. However, it is interesting to note that *HAMB* also provided ranges of concentrations for some of these known associations; *i.e.* although the association itself is not novel, the boundaries of the numerical values involved in the association may provide significant new information. A categorized subset of over 300 of *HAMB*'s rules can be found in the supplementary material.

### 3.4. Evaluation of *HAMB*'s use of the additional knowledge

In order to substantiate our claim that *HAMB* effectively uses the chemical and crystal-growth knowledge provided to it, we performed a lesion study to evaluate the effectiveness of some of *HAMB*'s heuristics that use domain-specific knowledge. To perform this study, we removed portions of the knowledge given to *HAMB* or disabled the portions of *HAMB* that use the knowledge. We used 500 cases randomly selected from the augmented BMCD.

The unmodified version of *HAMB* with the accumulated crystal-growth knowledge was also run on this set of cases, as was a version of *HAMB* that used no domain-specific knowledge.

The types of knowledge used by *HAMB* that we tested are as follows.

(i) Synonyms. If an attribute found in the feature set is synonymous (as either discovered by *HAMB* or stated in the domain theory) with another attribute in the feature set, the

attribute with the lesser estimated interestingness is removed from the set of attributes used to form the discoveries. A baseline version of *HAMB* omitted this capability and did not eliminate synonyms. It allowed the creation of 40 (19%) more redundant rules than did *HAMB* with heuristics for dealing with synonyms. This was surprisingly low, because the data contain many similar attributes. However, *HAMB*'s definition of redundancy is very strict, requiring either intensional (stated in the knowledge given to *HAMB*) or extensional equivalence (identical values for the cases); therefore, only a few pairs of attributes met its strict criterion for similarity. For example, the additive sodium azide and the species azide are synonyms.

(ii) Uninteresting attributes or values. The knowledge given to *HAMB* may contain information about attributes and values that are uninteresting or meaningless to the user. *HAMB*'s heuristics use this knowledge to avoid inducing rules containing uninteresting features (either in the left-hand side or right-hand side of a rule). A version of *HAMB* with capability omitted allowed the generation of 300 (141%) additional uninteresting rules. Examples of such rules involve attributes whose values could be 'not applicable', 'unknown' or 'misc'. Since *HAMB* generates and tests all possible attribute value associations in the first iteration, rules that contain uninteresting values for attributes on the left- or right-hand side are dropped from consideration during the next iteration. An example would be the broad association between a chemical additive with unknown role and the class of all proteins. By dropping such rules from further consideration, *HAMB* can focus on the more interesting attributes and rules that offer more information to the user.

(iii) Known associations. *HAMB* uses some of the knowledge given to it to remove attributes that have a known association (by causation, definition, association *etc*.) with the current target attribute. *HAMB* also removes attributes that are discovered to be extensionally equivalent to the target attribute. The version of *HAMB* used to test these heuristics omitted this use of knowledge and allowed the generation of 2897 (1367%) additional non-novel rules. Rules that represent rediscoveries that are well known in the 'lore' fall into the non-novel category. Examples include associations between nucleic acid class of macromolecules and the presence of cacodylate buffers and magnesium chloride and the protein class with the absence of magnesium chloride. This example depicts the common choices of cacodylate buffers and magnesium chloride in crystallization of nucleic acids. Also, those discoveries that have clear chemical explanations are also non-novel. For example, the association between the presence of a chelator and presence of highly charged species is non-novel, since EDTA is a highly charged chelator in common use.

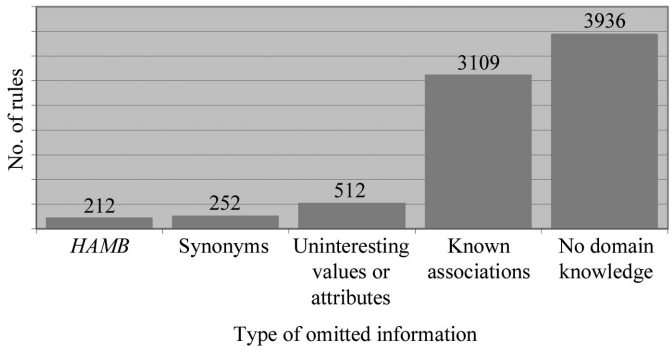


**Figure 5**
Graph of the number of rules reported *versus* the type of omitted domain information. The baseline version of *HAMB* reported 212 rules. Disabling *HAMB*'s ability to use knowledge about synonym attributes (given to *HAMB* or discovered) allowed *HAMB* to report an additional 40 redundant rules. Omitting *HAMB*'s knowledge about uninteresting attributes or values allowed *HAMB* to report an additional 300 rules, which were, by definition, uninteresting. Omitting knowledge about known associations among the attributes allowed *HAMB* to report an additional 2897 non-novel rules. The version of *HAMB* that used no domain knowledge reported an additional 3724 redundant, uninteresting or non-novel rules.

The regular version of *HAMB* reported 212 rules in this experiment, whereas the baseline version that used no domain knowledge reported 3936 rules. Thus, *HAMB* was able to use chemical and crystal-growth knowledge to avoid the creation of 3724 uninteresting rules. While the number of interesting rules is about the same in each case, the number of uninteresting rules shown to the user is much lower when using the regular version of *HAMB*, causing the percentage of interesting rules shown to the user to be much higher. Fig. 5 shows a graph of the results of applying the different types of domain knowledge constraints using *HAMB* on the augmented BMCD database.

### 3.5. Strengths and weaknesses of the BMCD *versus* strengths and weaknesses of *RL*/*HAMB*

There is no question as to the value of using experimental conditions from previous successful crystallizations as a guide to the design of new trials. However, as noted by Jurisica *et al.* (2001), there are two important limitations to the data provided in the BMCD. Firstly, it only includes information on the final successful attempt to grow a crystal. Information about the exact 'process' by which this final set of crystallization conditions were arrived at is unavailable. Secondly, many of the entries in the database are missing values for a significant number of important fields, such as the list of chemical additives used in the experiment.

Our approach involves: (i) the augmentation of BMCD with richer descriptions such as hierarchies of macromolecules (Hennessy *et al*., 2000) and chemical additives (described in this paper), (ii) the definition and encoding of relationships within the crystallization domain by attaching measures of interestingness to their examination within rule-based models and (iii) applying machine-learning tools such as *RL* and *HAMB* to make interesting discoveries from this augmented data. Owing to the fact that the impact of the learned discoveries is directly dependent on both the data as well as the heuristics that encode relationships within the domain, our algorithms are biased toward reporting discoveries that have the most evidence within the data and are of interest to the investigator. For example, we find the distribution of pH

values in the BMCD for ligand-binding proteins is significantly different from those reported for enzymes 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 proteins and enzymes show the opposite behavior, in which 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.

These results from our analyses provide objective support for both the idea that there are patterns of crystallization within the BMCD data, as well as the idea that classification schemes such as those we have developed capture some of these patterns. Nevertheless, there is a caveat with the use of techniques such as the *RL*/*HAMB* approach. The kinds of patterns highlight and reflect those areas of the parameter space that have abundant data. Therefore, by adding hierarchies to certain attributes, we bias the outcome of machine-learning and statistical approaches to consider patterns that mostly involve those attributes.

## 4. Conclusions

One of the challenges in automating the design of crystallization experiments is the representation of appropriate elements of the chemical knowledge people bring to this effort. Lists of additives (or the names of proteins) are simply text strings to the computer. However, the results shown here suggest that hierarchical classification schemes, such those presented here for 'commonly perceived role' (Fig. 2) do systematize important elements of the beliefs that have gone into the design of many crystallization experiments.

Another challenge is to automatically find the associations present in collections of this type of data. The results presented here also show that 'rule-learning' algorithms such as *RL* can detect many associations present in the data and represent them as if–then rules of the generic type if ⟨condition⟩ then ⟨outcome⟩.

However, rule-learning algorithms have a serious drawback: their outputs are voluminous and in this application the majority of the rules generated simply restated the knowledge represented explicitly or implicitly in the hierarchical classification schemes, *i.e.* they were of little use in efforts to grow better crystals. Therefore, a third challenge is to 'filter' the rules and to detect patterns in groups of rules to reduce the output to a manageable set of useful rule families.

The heuristically based *HAMB* algorithm described here addresses that challenge. The heuristics themselves represent a further systematization of the belief system used in crystal growth that ranges from restatements of the chemical knowledge to estimates of the potential interest of a particular set of associations. Heuristics of the former type filter out trivial rules that would not enhance our ability to grow crystals, while those of the latter guide the entire automated process towards the most useful set of rule families.

The results presented here range from the kind of things a well trained graduate student would be expected to know to potentially novel discoveries. It is noteworthy that these associations were found by the algorithms based on the hierarchies and heuristics described here. An example of the 'confidence-building' discoveries is the association between the crystallized macromolecule being a nucleic acid and the presence of $Mg^{2+}$ ions. It is also noteworthy that in addition to finding the magnesium–nucleic acid association, *RL*/*HAMB* also established an upper bound on the effective magnesium concentration (40 m*M*). An example of a potentially novel discovery is the rule that proteins containing a heme (or heme-like) prosthetic group crystallize at high ionic strength.

This research was 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 Supercomputer Center, NIH National Center for Research Resources (NCRR) grant NIHRR10447, the National Institute of General Medical Sciences grant GM62221 and National Library of Medicine Grant 2T15LM\DE07059. We gratefully acknowledge Devika Subramaniam for stimulating discussions and encouragement during the early phases of this project.

## References

Clearwater, S. & Provost, F. (1990). In *Proceedings of the Second International IEEE Conference on Tools for Artificial Intelligence (TAI-90)*. Los Alamitos, CA, USA: IEEE.

Feigenbaum, E. A. & Buchanan, B. G. (1993). *Artif. Intell.* **59**, 233–240.

Fisher, D. H. (1987). *Mach. Learn.* **2**, 139–172.

Gilliland, G. L. (1988). *J. Cryst. Growth*, **90**, 51–59.

Gopalakrishnan, V., Hennessy, D., Buchanan, B., Subramanian, D., Wilcosz, P. A., Chandrasekhar, K. & Rosenberg, J. M. (1994). *Preliminary Tests of Machine Learning Tools for the Analysis of Biological Macromolecular Crystallization Data.* Technical Report ISL-94-17. Department of Computer Science, University of Pittsburgh, USA.

Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P. A. & Rosenberg, J. M. (2000). *Acta Cryst.* D**56**, 817–827.

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, p. 179–187. Menlo Park, CA, USA: AAAI Press.

Jurisica, I., Rogers, P., Glasgow, J. I., Fortier, S., Luft, J. R., Wolfley, J. R., Bianca, M. A., Weeks, D. R. & DeTitta, G. T. (2001). *IBM Syst. J.* **40**, 394–409.

Livingston, G., Rosenberg, J. M. & Buchanan, B. (2001). *Proceedings of the IEEE International Conference on Data Mining, 2001*, pp. 385–392. Los Alamitos, CA, USA: IEEE.

Livingston, G., Rosenberg, J. M. & Buchanan, B. (2001). *Proceedings of the IEEE International Conference on Data Mining, 2001*, pp. 393–400. Los Alamitos, CA, USA: IEEE.

Mitchell, T. (1997). *Machine Learning.* Columbus, OH, USA: McGraw–Hill.

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

Sokal, R. R. & Rohlf, F. J. (1969). *Biometry: The Principles and Practice of Statistics in Biological Research.* San Fransisco, CA, USA: W. H. Freeman & Co.