Using Machine Learning to Analyze Biological Macromolecular Crystallization Data

Vanathi Gopalakrishnan, Daniel Hennessy, Bruce Buchanan*
{vanathi, hennessy, buchanan}@cs.pitt.edu
Intelligent Systems Laboratory
University of Pittsburgh
Pittsburgh, PA 15260

John Rosenberg
jmr@jmr3.xtal.pitt.edu
Departments of Biological Sciences and Crystallography
University of Pittsburgh

Devika Subramanian
devika@cs.cornell.edu
Computer Science Department
Cornell University

Abstract

The crystallization of a new macromolecule is still very much a trial and error process. In an effort to uncover useful trends in the crystallization of new macromolecules, Samudzi, Fivash and Rosenberg[12] performed a cluster analysis on the Biological Macromolecule Crystallization Database(BMCD)[7]. The crystallization parameters that were studied in order to differentiate among the experiments were a subset of the BMCD parameters: pH, temperature, molecular weight, macromolecular concentration, precipitant type and crystallization method. Samudzi et al. performed a purely statistical analysis of the data, and reported the clusters by eye-balling the results. We have attempted to recreate their clusters using two different methods - SAS clustering (same as Samudzi's) and COBWEB (a machine learning and discovery program). We then applied RL, an inductive learning program, to the discovered clusters from each of the methods, and verified as well as expanded on the Samudzi results. Apart from using clusters as the data input to RL, we also used RL on the entire BMCD data in an attempt to learn interesting correlations among the various crystallization parameters. From the point of view of crystallography, we have discovered possibly significant new empirical relationships. From a machine learning perspective, our work has led to the refinement of existing methods for incorporating detailed domain knowledge into inductive analysis techniques. In this paper we report these initial experiments and findings from applying RL to the BMCD as well as the Samudzi and COBWEB clusters.

*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.

1 Introduction

1.1 Background and rationale

Crystallization is the essential first step in macromolecular structure determination by X-ray crystallography. This is the only method capable of revealing high resolution structures for most proteins, protein-DNA complexes, viruses etc. The high resolution structural information is critical for modern molecular biological methods which rely on knowledge of the geometrical interrelationships of the various components that comprise the overall structure. (Multidimensional NMR can also determine macromolecular structures, but it is limited because it cannot be applied to molecules whose molecular weight is over 20,000; most proteins are larger than that).

The rate limiting step in X-ray structure determination is the crystallization itself. It takes anywhere between a few weeks to several years to obtain macromolecular crystals that yield good diffraction patterns. The theory of forces that promote and maintain crystal growth is preliminary, and crystallographers systematically search a large parameter space of experimental settings to grow good crystals. Thus there is a wealth of experimental data representing successful as well as failed crystallization trials.

Our goal is to use tools in machine learning to explore and identify patterns in the sizable historical data on crystallization trials. In particular, we will induce theories that capture relationships (correlations as well as causality) between experimental parameters, specific experimental protocols as well as protein characteristics that determine whether or not a diffraction-quality crystal is obtained. Such empirically derived theories can improve the probability of success of future crystallization trials on new proteins and contribute to the development of theories of processes underlying crystal growth.

Gilliland[7] has constructed a registry called the Biological Macromolecule Crystallization Database (BMCD), which consists of a large collection of crystallization trials, assembled over several years by crystallographers. Samudzi, Fivash and Rosenberg [12] have performed statistical clustering analysis on the BMCD data which revealed interesting qualitative relationships between recorded parameters. In particular, they demonstrated that rather than the data representing a single coherent domain, it more accurately represented a *set* of disjoint domains. However, very little information about the nature of those relationships or predictive information about how they could be used in the design of experiments could be obtained using the statistical techniques.

In this paper, we report experiments done with the BMCD data, which confirm and extend the analysis performed by Samudzi et al. Our approach is based on some initial experiments with a machine induction tool called RL [5]. We describe the unique features of the data domain that makes it a challenging induction problem and the aspects of RL that makes it ideal as a data exploration tool. We recreated the Samudzi clusters using both the SAS (FASTCLUS)[13] technique, as well as by using a popular discovery program called COBWEB. We then used RL to learn the correlations that existed among experimental parameters, within each cluster. This exercise confirmed the cluster descriptions that were described in the Samudzi paper, as well as exposed additional meaningful relationships between parameters.

1.2 Overall objectives

Our work is part of a larger project, called the Crystallographer's Assistant, a joint effort of the Intelligent Systems Laboratory and the Crystallography Laboratories at the University of Pittsburgh. Our overall objective is twofold (1) to provide computational tools to crystallographers to analyze the massive data sets of crystallization trials, as well as assist in the retrieval and design of experimental protocols (2) to develop more powerful knowledge-based induction tools with automatic reformulation abilities and to design an interactive assistant for experiment planning and simulation of processes in crystal growth.

1.3 The BMCD

The Biological Macromolecule Crystallization Database (BMCD), consists of 1025 *successful* crystallization conditions of over 820 macromolecules. Data for proteins, protein-protein complexes, nucleic acids, nucleic-acid:nucleic-acid complexes, protein:nucleic-acid complexes, as well as viruses have been recorded. Data about the macromolecule include names and class names, molecular weight, source tissue, and presence of prosthetic groups. Crystal parameters are also recorded. Experimental conditions include pH, temperature, growth time, crystallization method, macromolecular concentration and chemical additives to the growth medium (providing both symbolic as well as numeric data).

The creation of the BMCD is an important first step toward progress on systematic data collection in the field. The BMCD is not an on-line database — it is updated periodically by its creators. It is simply a record of successful experimental conditions. In addition, the commercial version offers limited tools for analysis or design.

There are two important limitations to the data provided in the BMCD. First, it only includes information on the final, successful attempt to grow a crystal. Information about the typically vast number of experiments which failed to grow adequate crystals (or any crystals at all) is not available. Second, many of the entries in the database are missing values for a significant number of their fields. For instance, the complete list of chemical additives, an important factor in the ability to successfully grow a crystal, is not reported for most of the entries of the database.

In order to be able to use our tools to perform experiments with the BMCD data, there was a need for extensive modification of the existing format of the data.

2 The Samudzi paper - insights and problems

2.1 Cluster Analysis

In order to form clusters or groups of similar objects, some measures of differences or similarities must be available. Samudzi, Fivash and Rosenberg have performed statistical clustering analysis on the BMCD data, by defining a method of measuring the differences between experiments, and then employing the PROC FASTCLUS method to locate clusters in the BMCD using this Euclidean metric.

Each crystallization experiment recorded in the BMCD is described by a set of parameters, some of which describe a continuum, while others describe a category. The overall measure of difference between two crystallizations was formed by the construction of an n-dimensional space, one dimension for each continuous parameter and m_i - 1 dimensions for the i^{th} categorical parameter.

The problem with the above construction of measure of difference is in the computational complexity. With each additional value of a categorical parameter, the complexity of the computational space increases. The Samudzi experiment was described by a vector dimension of forty-four. [Give an example of the computational complexity involved -state that of the 6 parameters, 4 were continuous and only 2 were categorical, and give a numerical example of computational complexity.] [Do we need to mention the Pseudo-F statistic and the 3 criteria used to evaluate the propriety of the number of clusters - Samudzi et al. got 6, 7, 8 9 and 10 clusters - and then decided to go with 8 based on these 3 criteria? How does this tie in with our approach?]

Parameter values were normalized, and parameters were weighted to determine cluster stability. Their experiments with weighting of parameters indicated that the clusters were stable, as the results from "weighting" did not show significant differences in the way the clusters were formed, as compared to equal weighting.

The results of the above cluster analysis were then analyzed by eye-balling the clusters, to yield cluster characterizations which revealed interesting qualitative relationships between recorded parameters. However, very little information about the nature of those relationships or predictive information about how they could be used in the design of experiments could be obtained using the statistical techniques.

3 Extensions to the Samudzi work

Our main goals in attempting to recreate the Samudzi clusters were to:

1. verify the accuracy of the clusters - by using two different methods to obtain clusters of the BMCD data;

2. run RL on each of the clusters to learn significant relationships among the crystallization parameters within each cluster and hence to

3. match the output of RL against the cluster characterizations obtained by eye-balling the data, and identify the differences and similarities. This would also enable us to evaluate RL's output in this domain. Unlike statistical clustering analysis, RL has the added advantages of being able to (1) exploit and analyze symbolic data (as well as automatically select the most effective boundaries for numeric data) and (2) exploit prior knowledge of the domain to guide the analysis. Furthermore, RL's ability to deal with noise and explicitly represent and systematically manipulate its bias make it an ideal tool for this domain. As such, we expected to confirm as well as extend and explain in greater detail the results of the previous work.

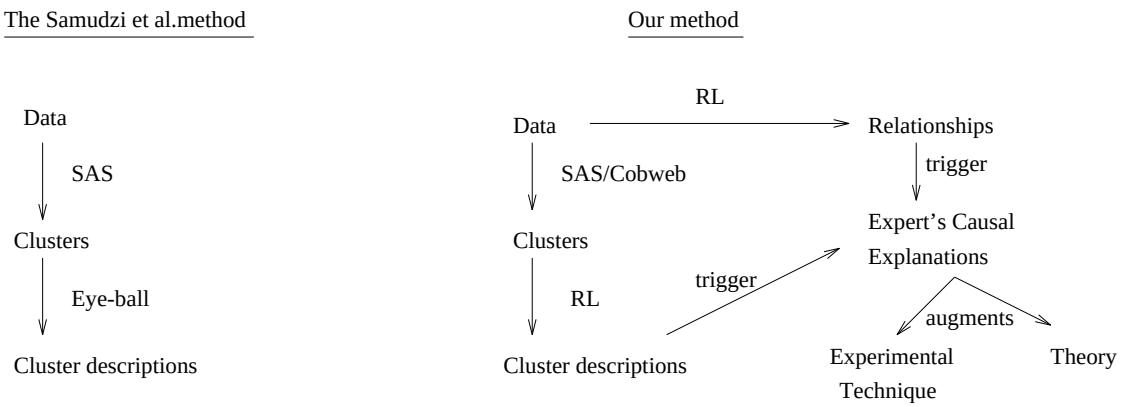

Figure 1: A comparison of the two approaches.

Figure 1 shows the comparison between our approach and that of Samudzi et al. Two different clustering methods were used in order to verify the Samudzi clusters - SAS and COBWEB. (Cobweb is a machine learning and discovery program that uses a top-down approach to build the clusters of instances.) The clusters we get using both methods are very similar to the Samudzi clusters, except that we get a larger number of clusters. Moreover, instead of eyeballing the data to describe the features of the clusters, we employed RL to learn the correlations in the data. This yielded a super-set of the relationships described by Samudzi et al.

We were able to:

- recreate the Samudzi clusters accurately with respect to the numbers of different types of data that were specified in the Samudzi tables.
- get an extended set of clusters using the two methods described in the following section.
- derive more accurate statistical relationships among data within each cluster by using machine learning.

4 Machine Learning Methods

4.1 Inductive Learning

Machine learning from examples, or inductive learning, has received considerable attention since the 1950's, with several approaches now in the growing tool kit of knowledge-based programs that have been demonstrated to work[3]. We are using a machine learning program developed in our laboratory, named RL, to explore data from a database of crystallography results (see the section on the BMCD). We believe that RL has found rules that characterize some of the replicable conditions that increase the chances of success.

RL[5] is a knowledge-based inductive rule learning program, that induces one or more **IF**-*condition*-**THEN**-*class* rules from specific examples of classes.[1]

[1]An exportable version of RL is available on request. The exportable versions of RL are written in

For example, in our study, we partition the BMCD data based on the diffraction limit of a crystal - e.g., diflim < 3 and diflim > 3Å. RL is given a set of BMCD data that is partitioned based on the diffraction limit, and it induces one or more rules which describe correlations between experimental parameter values and the diffraction of crystals obtained with those conditions. RL uses a heuristic rule search which can utilize prior domain knowledge, such as facts or heuristics used by scientists, and can examine a much larger number of identification criteria than can be examined with manual analysis.

The main strength of RL is its flexibility. Given a learning problem, many different problem models and assumptions can be tested. The flexibility is partly achieved through the use of a domain model, called the Partial Domain Model (PDM), which can guide RL's search separately from the guidance implicit in the statistics of training examples. The PDM contains definitions of attributes to be used in representing examples and rules, a list of classes, assumptions and constraints on rules being sought, and domain knowledge relevant to a particular problem. Constraints and domain knowledge usually take the form of preference criteria characterizing desirable properties of rules to be learned. Thus, induction in RL is guided not only by syntactic similarity and dissimilarity of features of examples, but also by constraints and prior domain knowledge in the PDM.

For RL, an example is represented as a vector of attribute-value pairs, each of which describes a feature of the example. For example, the representation of a feature subset of an entry or experiment in the BMCD is shown in Table 1.

Representation	((macmol "Alcohol-dehydrogenase") (maccon 5) (crmethod Bulk-Dialysis) (srctis liver) (srcgsp equus-caballus) (pH 8.4) (temp 4.0) (buffer Tris) (spacgp C222-1) (diflim 2.4))
Meaning	An entry in the BMCD that contains the successful crystallization conditions of the macromolecule Alcohol-dehydrogenase obtained from the liver of a horse, used the bulk-dialysis method of crystallization. The favorable experimental conditions were a macromolecular concentration of 5 mg/ml, pH of 8.4, the buffer Tris, and temperature of 4^oC. The crystal that resulted from this experiment diffracted well with a diflim of 2.4Å, and belonged to the space-group C222-1.

Table 1: Attribute-value pair representation of a subset of features in an experiment in the BMCD database.

Given a learning problem, *i.e.*, the names of one or more target classes, a set of their examples, and a PDM of the problem, RL searches for rules by examining a large but limited number of combinations of features. In most learning problems, RL prefers a general rule, i.e., a rule classifying more training examples correctly, to a specific rule. However, RL can search for more specific rules to the extent specified in the PDM. Only those rules which are plausible (with respect to the PDM and training examples) are saved in the

CommonLISP and C/C++ and run on any UNIX workstation with more than 4MB of memory. However, the export versions do not include new features which are still under development, and some of which were used in the present study.

resulting set of rules. That is, the plausibility of a rule is determined by its performance (how accurately it classifies examples) and its concordance with assumptions, constraints, and domain knowledge.

The result of rule search is a disjunction of **IF**-*condition*-**THEN**-*class* rules, where condition is a conjunction ("AND") of features. For example, the following rule uses two features to predict crystal diffraction quality:

IF-(buffer cacodylate) and (molwgt $>$ 39630)-**THEN**-(DIFLIM-UNDER-3)

which is interpreted as *if the buffer cacodylate is used to crystallize macromolecules that have a molecular weight over 39,630 kilo Daltons; then you can obtain well-diffracting crystals.* Each rule has associated with it a statistically computed number, which indicates the likelihood of accuracy of the prediction. This will be explained in detail in a later section. Such **IF-THEN** rules are very easy to understand, unlike numerical weights and neurons in a neural network. The comprehensibility of rules facilitates the easy verification of rules by experts.

Some additional features of RL (which can also be found separately in other methods) are: It can

- exploit redundancy in the attributes (rather than insisting that they be independent).
- integrate numeric and symbolic attributes. Semantic relationships among symbolic attributes can be exploited. Meaningful intervals of continuous numeric attributes are learned automatically.
- tolerate erroneous and missing data. The training data may be presented to the program incrementally, that is they do not all need to be in the database at once.
- use prior knowledge about the relative costs of false positive and false negative predictions to learn more appropriate rules. RL learns rules that make their predictions with different degrees of strength.

RL is a descendant of the Meta-DENDRAL system [6], which specialized in finding rules of mass spectrometry in chemistry. However, unlike the Meta-DENDRAL, RL is a general purpose learning program that can be applied to many problems in different domains. RL has been applied successfully in several real world problems, including learning predictive rules for carcinogenic activity in chemicals [1], learning filtering criteria for high energy physics experiments [10], and learning diagnostic rules for telephone line problems [11].

4.2 Conceptual Clustering

The RL program *learns by examples.* In contrast, COBWEB is a machine learning and discovery program that does *conceptual clustering*[9]. The main difference between programs that perform learning and those that perform discovery is that in the former, a pre-classified set of examples about a particular concept are given as input to the program and the program learns combinations of features that constitute a description of the concept being learned. In the latter, the examples are not pre-classified. That is, COBWEB does not require predefined

class information in its input. Another difference between COBWEB and other machine learning systems is that instead of forming flat classes that merely group similar instances together as do iterative optimization algorithms[2][4], COBWEB forms a *concept hierarchy*. In the hierarchy, nodes represent concepts that are partially ordered by generality. Each node includes a concept description, a list of probabilities that certain features will occur in instances covered by the concept. Figure 2 shows the hierarchical organization.

Conceptual clustering systems accept object descriptions and produce a classification scheme over the observed objects. COBWEB transforms a collection of object descriptions into a classification tree, where objects are described in terms of nominal attribute-value pairs as shown in Figure 2.

[Use figure to show the COBWEB tree with three children containing BMCD instances]

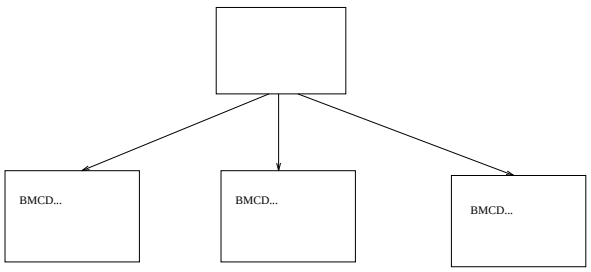

Figure 2: COBWEB hierarchy of BMCD instances.

Each BMCD instance was given a name: BMCD<#>, where # represented a number between 1 and 1025, i.e. the sequence number of the particular BMCD instance.

5 Statistical Clustering

[Devika's description]

6 Experiments with the BMCD data

In order to analyze the BMCD data, one of the initial tasks was to convert it into a form usable by RL. Most of this data conversion could be automated, although some of the fields required significant reengineering. This reengineering took on three forms: data re-representation, attribute abstraction and data subsetting. Many of the fields in the BMCD (both numeric and symbolic) required re-representing or normalization. Many of the symbolic fields used multiple terms for the same concept while some of the numeric attributes used multiple scales for a single feature. Furthermore, some attributes represented very complex features which were more readily usable as a set of attributes. This process of re-representing the data was important in creating a consistent set of data represented at a level that would provide maximum utility to RL.

Attribute abstraction, the generalization of feature values into hierarchies, was also used to increase the power and flexibility of the data. For instance, there were approximately 280 chemical additives amongst the entries in the database. These additives however, had

not been classified in the BMCD database into different classes such as precipitating agent, buffer, etc. As a result of an inter-disciplinary effort that included experts in crystallography, we were able to hierarchically organize the additives according to their common use, and then convert the BMCD data based on the classification. RL has the ability to utilize these attribute abstractions to help guide the specialization of rules.

A similar issue arose in the classification of the examples. As mentioned earlier the BMCD contains only positive instances (i.e. successful crystallization attempts) while RL, as it is a supervised learner, needs both positive and negative instances to be able to generate rules. We therefore decided to use the diffraction limit to define "success" and "failure" with crystals diffracting to resolutions higher than the limit being defined as acceptable (it should be noted that many of the BMCD entries diffract to relatively low resolutions, e.g., 7.0$\mathring{A}$). We initially used a threshold of 3.5$\mathring{A}$ based upon the fact that it is difficult (almost impossible) to determine the structure for resolutions above that value, but later choose more stringent values of 3.0$\mathring{A}$ and 2.5$\mathring{A}$ to strengthen the rules.

Data-subsetting is the selection of portions of the data in order to reduce the amount of irrelevant or noisy data. This can take on two forms, subsetting of the attributes and subsetting of the examples. To date, we have focused on the selection of subsets of the attributes. For instance, we often ignore the attributes which have a coverage for less than 40% of the entries in the database. This selection criterion is the same as that used by Samudzi et al. Also, in order to obtain more specialized relationships from the data, we conducted experiments which use only subsets of the BMCD entries (e.g., only protein crystallizations, individual clusters, etc).

New meaningful attributes were added to enhance the readability of the rules that were output by RL. For instance, an individual spacegroup was converted to represent four different attributes, viz. *crystal system*, *point group*, *centering* and *polymorphic*[2]. The crystal system values were based on the Bravais Lattice. By reorganizing the data about spacegroups in this manner, we were able to obtain additional relationships between the input and output parameter values of a crystallization experiment.

6.1 Hypotheses choice and experiment design

Two of the parameters that can be tuned to improve RL's performance are the positive and negative coverage thresholds for "good" rules. These are important because they define the bias with which RL looks at the data and how it is allowed to deal with noise in the data. A high positive threshold will force the search towards more general rules, but rules which do a good job of covering a small set of examples (which may or may not be due to noise) will not be explored. However, if the positive threshold is too low, large numbers of rules originating from coincidental correlations in the data will be generated. Similarly, a high negative threshold will generate a large number of general rules, while a low negative threshold, while forcing RL to further specialize its results, can prune possibly interesting rules or force it to overfit the data.

In these experiments we used small values for both the positive and negative thresholds.

[2] *Polymorphic* is an attribute that takes as values YES or NO. It indicates whether there were several space groups that were recorded in one single entry in the BMCD data or not.

This generates a large number of rules each of which may not cover very many of the positive examples, but will only include a very small percentage of the negative examples. This combination of thresholds was chosen because, as mentioned above, we were interested in discovering interesting relationships over the smaller disjoint sets suggested by Samudzi's results, not necessarily creating a set of rules which cover the entire set of positive examples. A low positive threshold generates highly specialized rules.

Our initial version of RL did not provide results that were sorted based on the certainty factor (defined below) of rules. We modified RL to sort the generated rules based upon certainty factor, and the number of positive examples covered. The new version of RL orders the rules such that rules with the highest coverage are output first. This was helpful in identifying the significant results.

6.2 Criteria for experiment design

There are three types of parameters in the BMCD: (i) *givens* (molecular weight, macromolecular class, etc.) (ii) *controllables* (pH, temperature, macromolecular concentration, chemical additives, buffers, crystallization method, etc.), and (iii) *observables* (crystal habit, polymorphism, diffraction limit, etc.). Rules can relate some or all of the parameters of the these types to one another. The most useful rules from the perspective of designing new experimental trials relate *givens* to *controllables* and *observables*.

6.3 SAS experiments

The PROC FASTCLUS method used by Samudzi et al. was also used by us to recreate the Samudzi clusters. The clusters were created accurately with respect to the number of the different types of BMCD instances that are recorded in the tables in the Samudzi paper. Much of the difficulty in reproducing the results was because of the massive amount of data re-organization that had to performed on the values of the 6 parameters in the BMCD. Some of the important data re-organization tasks were:

- Macromolecule concentration: The values for this parameter in the BMCD data varied depending on the units, viz. mg/ml, or number of moles/liter. All the values had to be converted to values in one unit.

- Values of Molecular weight, pH, preicipitant type, and temperature had to be normalized. [Devika, Dan - needs to be expanded...]

RL was used on this data and it found a super-set of the relationships within each cluster that were described by Samudzi et al. Apart from confirming the cluster descriptions, RL found further meaningful relationships to add to each cluster description. The RL experiments with the SAS cluters were designed to find the the set of all possible relationships that existed between the entries in each cluster. The main RL parameters that were varied in order to achieve this goal were the *certainty factor* thresholds and rule coverage. The results of the experiments will be described in the following section.

6.4 COBWEB experiments

We converted the BMCD data so that it can be input into COBWEB. We extracted the COBWEB clusters at two different levels - one which is an immediate clustering of the BMCD data (3 clusters were formed) and another which was a further division of these 3 clusters - Cluster 1 had two children, Cluster 2 had three, and cluster 3 had 4; bringing the total upto 9 clusters at this level. The clusters discovered by COBWEB resemble the Samudzi clusters closely in terms of the descriptions obtained from RL for each cluster. The results are described in the following section.

7 Results

Table 2 compares the results obtained from the various sets of experiments described in this paper. The table shows the compression factor, the total number of rules, and the number of interesting rules produced by RL for three settings of the diffraction limit. The interesting rules are classified into six categories viz., *givens* $\rightarrow$ *givens*, *controllables* $\rightarrow$ *controllables*, *observables* $\rightarrow$ *observables*, *givens* $\rightarrow$ *controllables*, *givens* $\rightarrow$ *observables*, and *observables* $\rightarrow$ *controllables*. Some of the other RL runs yielded rules that connected *givens* to both *observables* and *controllables*. As the diffraction limit increases, the total number of rules induced by RL increases because the number of negative instances in the database is reduced. Note that several of the interesting rules (Types IV, V and VI) form the bases of a predictive theory of crystal growth. We cannot judge rule sets merely by the number of interesting rules contained in them. For instance, there are 31 interesting rules produced with a diffraction limit of $3.5\mathring{A}$ while only 21 are generated with a limit of $3\mathring{A}$. However, the rules associated with a limit of $3\mathring{A}$ are judged to be more useful by the crystallographers due to the fact that crystals with higher diffraction limits do not produce good structure maps. A remaining challenge in this domain is to devise more objective criteria to capture the subjective rules used by our expert in judging the quality of rule sets as well as individual rules. The impact of both attribute as well as data-subsetting can be evaluated from the table and the following graph.

Figure 3 summarizes the evolution of the quality of rule sets as we enriched the BMCD data set by data re-representation, attribute abstraction and data subsetting. We use the number of interesting rules identified by our expert as the quality metric here. A similar graph can be generated for other metrics of rule set quality. By standardizing the symbolic names in the data set and running RL on the raw BMCD data we observed a sizable increase in the number and the quality of the interesting rules. The reformulation of the *chemad* attribute into 280 individual attributes to handle the set-valued nature of *chemad* resulted in a significant drop in rule set quality. By classifying these attributes according to their functional roles, we were able to restore and in fact, improve, the overall quality of the rule set. It became clear that the BMCD was not a homogeneous database and that we needed to partition the data according to the type of macromolecule to obtain good rule sets. We isolated 545 out of the 1025 cases pertaining to proteins for subsequent rounds of experiments. Further improvements required us to focus on specific attribute sets identified by our expert to be "relevant" to crystal growth. This step generated the rules that excited

Diflim	Coverage	Total #rules	#interesting rules	#I	#II	#III	#IV	#V	#VI
2.5	313/545	104	25	3	0	3	10	4	5
3	361/545	112	21	4	0	5	3	0	9
3.5	334/545	199	31	3	2	8	5	3	10

I = givens to givens II = controllables to controllables III = observables to observables
IV = givens to controllables V = givens to observables VI = observables to controllables

Table 2: Summary of results from RL experiments with a subset of the BMCD data

our expert; some of them are presented in the next section. The hierarchical restructuring of attribute values for space groups resulted in general rules of high quality.

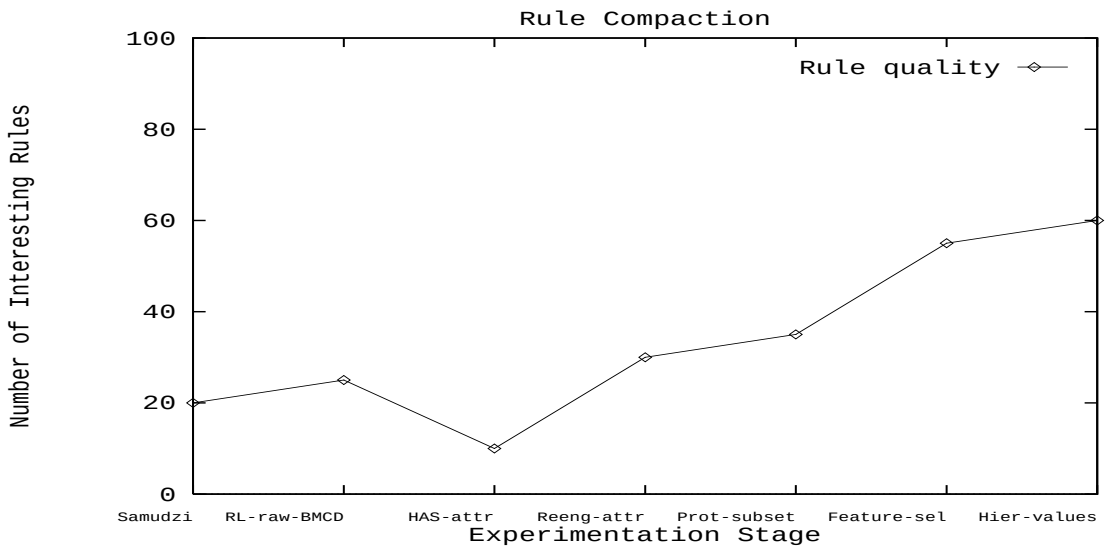

Figure 3: Rule set quality at each stage of experimentation

8 A Sampler of Rules discovered by RL

We have processed the 1025 entries in the BMCD through several RL iterations. This has resulted in the production of rules of the form:

> If the crystal habit had the value *plates* then the diffraction limit was under 3.5$\mathring{A}$ (i.e., the crystal diffracted moderately well).

(CRHABI PLATES) ==> DIFLIM-UNDER-3.5
pos= 0.09, neg= 0.05, cf= 0.890
p=45, n=5, tp=520, tn=109

pos and *neg* represent the percentages of positive and negative examples matched, p and n are the number of positive and negative examples in which the rule applies, *tp* and *tn* are the total number of true positive and true negative examples available, and certainty factor, *cf* is the number of positive examples matched over the total number of examples $p/(p+n)$ in which the rule applies. Please note that the antecedents of the rules can have multiple conjuncts.

Included in the output were: (a) rules that were consistent with the current conventional wisdom in crystallography, and (b) rules which suggested the possibility of useful relationships that were otherwise unknown in the research community, or that even appeared to contradict current "lore", suggesting the need for further study or clarification of the conditions under which these rules hold.

Interdisciplinary areas come into the fore here, because the crystallographers were able to construct interpretations for the rules that were generated by RL. For ease of understanding the rest of this paper presents the rules in the form of the crystallographer's interpretations.

8.1 Using RL on SAS clusters

[Table showing summary of results - comparing our SAS output to Samudzis]

8.2 Using RL on COWEB clusters

Table 3 summarizes the results from RL using the the two different levels of clusters obtained from the COBWEB experiments. There were no examples of Samudzi cluster 4 because of bias in the way we generated the data for the runs from the BMCD. COBWEB discovers the clusters that Samudzi et al. describe, but divides them in a different manner as shown in the table.

8.3 Using RL directly on BMCD data

8.3.1 Confirmation rules

The first group were *confidence building* rules that recognized patterns in the data that were expected, given theoretical and experimental results in crystallography. Some instances of such rules are:

- It is relatively easy to grow high quality crystals of lysozyme.
- Good crystals of lysozyme can be obtained over a wide range of pH.
- Similarly a few other proteins, including *cytochrome c*, *chymotrypsin* and *ribonuclease-t-1* also give good crystals.

Clusters type	Number of clusters	Description of clusters
RL-COBWEB-1	3	One cluster (C1) corresponds to Samudzi clusters (SC) 2, 5 and 8. Another cluster (C2) corresponds to half of SC 7 (179/329 entries covered). The third cluster (C3) corresponds to SC 1, 3, and the rest of 7 (150/329).
RL-COBWEB-2	9	These clusters depend are the children of C1, C2 and C3. C1 breaks down into 2 separate clusters: one corresponding to the entire SC 5 and the other containing both SC2 and SC8. C2 has three children, each containing a subset of SC 7. C3 has 4 children: (1) corresponds to SC 3 (76/91); (2) SC 7 and rest of SC 3; (3) all of SC 1; and (4) rest of SC 7.

Table 3: RL-COBWEB results

- Good crystals of small molecular weight proteins can be obtained relatively easily, while
- it is harder to obtain good crystals of very high molecular weight proteins.
- Macromolecules with more than one polypeptide chain (subunit) diffracted well (under $2.5\mathring{A}$) when the number of subunits was small (< 3).
- Large complicated systems diffract less well than smaller, simpler systems.
- Crystals with a rectangular prismatic "habit" (shape) tend to diffract well.

All of these rules are consistent with the general "lore" and experience in crystallography. Additionally, we found rules that result from "sociological causes". One example of this type of rule was that a combination of using blood as the source tissue and ammonium sulfate as an additive produced high quality crystals. It is most likely that this stems from an incidental correlation of the almost exclusive use of these items during the same period of time and the way the results were reported at that time.

A related example is the "rule" that synthetic oligonucleotides give good crystals. In fact, it is well known that synthetic oligonucleotides are extraordinarily difficult to crystallize. Workers in this area only report instances of diffraction to $3\mathring{A}$ resolution (or better), establishing a "sociological" pattern in the BMCD that was among the more consistent rules found by RL. Although this describes the behavior of people, rather than molecules, it shows that RL can find the patterns present in the data, which suggests that more complete data would yield rules more attuned to molecular behavior.

8.3.2 Discovered rules

The second class of rules were those that suggested possibly new relationships. One such observation was that orthorhombic space groups, especially $P2_12_12_1$, $C222_1$, and I222 tended

to be associated with well diffracting crystals. This could reflect a *generally* stable packing arrangement, and could be used as a guide in cases where small, polymorphic crystals are obtained and where one of those forms appeared to be orthorhombic. In that case, it would be reasonable to concentrate further efforts on that form, based upon the theory that orthorhombic forms have tended to give better diffraction.

- One interesting set of rules suggested that, contrary to conventional belief, the crystals which were needles and plates (as reported in the BMCD) tended to diffract well. (It should be noted that the "rule" that needles diffracted well was true only for small proteins. Similarly, the successes for plates generally tended to be in "simpler" systems with low molecular weight and/or no more than two subunits). This was not generally expected and one interpretation is sociological. Crystals that diffract well are generally more completely described in the literature. A well diffracting crystal in the form of a needle or plate would be noteworthy; hence information on the habit would probably be included in the published report. It is quite possible that many of the poorly diffracting crystals were also needles or plates, but this information was not included in the papers cited by Gilliland.

  We are also currently studying the possibility that these rules could actually be derivatives of the preceding rule *i.e.* that orthorhombic space groups also tend to produce needles and plates. This result also demonstrates the need for a more precise descriptive language: One persons "needle" may be another's "lath" and the relative thickness of the crystals is unclear. However, this rule could suggest that in experiments where it is difficult to grow anything other than needles or plates, it may still be valuable to continue the investigation, especially if the habit suggested that the space group might be orthorhombic.

- Another rule was that very high macromolecule concentrations (> 50 μg/ml) tended to have diffraction limits over $2.5\mathring{A}$. It is generally felt that high concentrations are better, but this suggests that there might be upper limits, perhaps due to aggregation. Other rules suggest that it may be possible to mitigate this problem with high pH ($>$ 9.0).

- A very interesting rule suggested the correlation between space group $P2_12_12_1$ and warmer temperatures for yielding crystals that diffract well. Based upon this correlation, a crystallographer might try to conduct a particular experiment at a warmer temperature, if it yielded a crystal with space group $P2_12_12_1$ without good resolution. There is no good explanation as to why $P2_12_12_1$ does not do so well when the experiment is conducted with cold temperatures, but the above rule could be used to test whether the experimental results improve at warmer temperatures.

- The choice of E. Coli as a prototypical procaryote may not have been the best for crystallographers. That rule had a large coverage of examples of E.Coli that yielded crystals that did not diffract well, and very few negative examples.

- Several rules included Gamma ($<= 90$) in cases where good crystals were obtained. This means the exclusion of Trigonal and Hexagonal space groups, which tends to

suggest that these groups will not diffract as well as the others.

- There is a recurring theme that Tris buffers do well. In fact, Tris does better than "Good" buffers[3]. While it might be thought that this could be masking a pH effect (Tris is commonly used near "physiological" pH values), it is interesting to note that "Good" buffers also include this pH range but they seem to not do well, especially for larger systems. Hence one message to crystallographers is "try Tris".
- There was also a weak rule that suggests that reducing agents are indeed helpful (when required).

Another feature of RL that has provided interesting results is its ability to automatically determine optimal boundaries or markers for numeric valued data. This produced an interesting class of rules that suggest bounds on the applicability of experimental conditions. For instance, it was found that:

- Sulfate does well in conjunction with pH values above 8.1.
- Alcohol as a precipitating agent does well for macromolecules with a molecular weight under 22,000 (there are very probably nucleic acids).
- Phosphate as a buffer does well with molecular weights between 12,920 and 110,000, while
- the buffer cacodylate seems to do well with molecular weights over 39,630.

The above rules represent relationships present in the data reported in the BMCD. As such, the accuracy and completeness of the rules is limited by the accuracy and completeness of the data. As discussed earlier, the BMCD suffers from major limitations in both of these areas. First, it only includes information on the final attempt to grow a crystal, excluding Information about failures. This means that we can make statements about those experimental conditions which might yield crystals, but not about those that might not. We are currently studying the dependence of our rules on the resolution cutoff. Second, many of the data are missing entries for a significant number of the fields in the database. This further limits the coverage, and therefore the accuracy, of the induced rules.

9 Future Work

The above preliminary work suggested a number of areas requiring deeper study in order to extend the results we have produced so far. These include:

1. Need for data on failed crystallization attempts, viz. data describing not just the success, but also the experimental conditions of the failures. This would allow more accurate and higher confidence rules to be inferred. In particular, it would give the rule learner important additional information, with which RL can judge when it begins to overgeneralize rules. We are currently collecting data of this type from one lab (Rosenberg) and will soon expand that to additional labs for use in future analysis.

[3] A series of buffers that were developed by Dr. Good[8]

2. Need for more complete data, including data on the sequence of experimental conditions attempted: crystallography experiments do not take place individually. Rather, a series of experiments are conducted where the results of one experiment are used to derive the experimental conditions for the next batch of experiments. This information would also be important in increasing the reliability and credibility of the induced rules. This experimental context is also being captured in the data that is currently being collected.

3. Need for use of knowledge-based inductive learning techniques. One of the primary advantages of machine-learning over statistical analysis techniques is its ability to incorporate existing domain knowledge into the analysis. The crystallography domain includes a significant corpus of domain knowledge about theoretical and experiential relationships that exist between concepts that relate to performing crystallography experiments. This knowledge can be represented both as a set of constraints on the types of relationships that RL could explore and as initial rules to seed induction. This would enhance the performance of RL in terms of speed, accuracy and reliability.

4. Need for the use of complex representations of the attributes. This is another form of domain knowledge that RL can use to exploit the relationships between symbolic values of the data (e.g., I222, $P2_12_12_1$, and $C222_1$ are all orthorhombic forms). Hierarchical representations of values would allow RL to produce powerful, more broadly applicable rules. Some information of this form is being exploited currently.

5. Need for a multi-level representation of the experiments. The experiments themselves can be represented at multiple levels, just as the above-mentioned value hierarchies. Representing the data in a fashion similar to the hierarchies of values, would allow RL to produce powerful, more broadly applicable rules.

6. Need for the joint participation of people with both machine learning and crystallographic experience. The former are clearly needed to use the techniques in their current form while the latter are essential to guide the development of the methods along crystallographically productive routes. One of the primary goals of this project is to produce programs that would allow a typical crystallographer to use these techniques independently.

10 Conclusions

One of the most encouraging aspects of using machine learning in the crystallography domain is the fact that, in spite of noisy and insufficient data, interesting correlations among attributes have been obtained. The results from our preliminary runs with RL using BMCD data are promising. Furthermore, RL has generated "better" rules than those generated by the statistical clustering analysis (in that they provide greater insight into the details of the discovered relationships) by including symbolic data and incorporating domain knowledge through attribute abstraction and data subsetting. From a machine learning point of view, our work has led to the refinement of methods for incorporating detailed domain knowledge into inductive analysis techniques.

Finally, the close relationship between the computer scientists and crystallographers cannot be overstressed. It has been critical to the analysis of the data, incorporation of domain knowledge, and the interpretation of the rules. This collaboration has resulted in significant refinements to our approach. We expect our continued collaboration to yield greater insight into the principles involved in the crystallization of biological macromolecules and the development of additional tools to enhance crystallographers' productivity through the Crystallographer's Assistant.

References

[1] Ambrosino, R., Lee, Y., Rosenkranz, H., Mattison, D., Provost, F., and Gomez, J. The use of a knowledge-based program to predict chemical carcinogenesis in rodents. In *Proceedings of the International Workshop of Predicting Chemical Carcinogenesis in Rodents* (1993).

[2] Anderberg, M. *Cluster Analysis for applications.* New York: Academic Press, 1973.

[3] Buchanan, B., and Wilkins, D. *Readings in Knowledge Acquisition and Learning: Automating Construction and Improvement of Expert Systems.* Morgan Kaufman, 1993.

[4] Cheeseman, P., and J. Kelly, e. AUTOCLASS: A Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning* (1988), Morgan Kaufmann.

[5] Clearwater, S., and Provost, F. RL4: A Tool for Knowledge-Based Induction. In *Proceedings of the Second International IEEE Conference on Tools for Artificial Intelligence* (1990), IEEE CS. Press, pp. 24–30.

[6] Feigenbaum, E., and Buchanan, B. DENDRAL and Meta-DENDRAL: roots of knowledge systems and expert system applications. *Artificial Intelligence Journal 59*, 1 (1993).

[7] Gilliland, G. C. A biological macromolecule crystallization database: a basis for a crystallization strategy. In *Crystal Growth of Biological Macromolecules*, R. Geigé, A. Ducruix, J. C. Fontecilla-Camps, R. S. Feigelson, R. Kern, and A. McPherson, Eds. North Holland, 1987. Proceedings of the Second International Conference on Protein Crystal Growth, Bischenberg, Strasbourg, France. A FEBS Advanced Lecture Course.

[8] Good, N., and G.D. Winget, e. Hydrogen ion buffers for biologica research. *Biochemistry 5* (1966), 467–477.

[9] Michalski, R., and Stepp, R. Learning from Observation: Conceptual clustering. In *Machine learning: An artificial intelligence approach.* San Mateo, CA: Morgan Kaufmann, 1983.

[10] Provost, F., and Buchanan, B. Inductive policy. In *Proceedings of AAAI-92* (1992).

[11] Provost, F., and Buchanan, B. Inductive strengthening: The effects of a simple heuristic for restricting hypothesis space search. In *Proceedings of the Third International Workshop on Analogical and Inductive Inference* (October 1992).

[12] Samudzi, C., Fivash, and Rosenberg, J. Cluster analysis of the biological macromolecular crystallization database. *Journal of Crystal Growth 123* (1992), 47–58.

[13] SAS. User's Guide: Statistics, Version 5 edition. Tech. rep., SAS Institute Inc. Cary, North Carolina, 1985.