Induction of Rules for Biological Macromolecule Crystallization

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

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

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

Abstract

X-ray crystallography is the method of choice for determining the 3-D structure of large macromolecules at a high enough resolution. The rate limiting step in 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. There is a wealth of experimental data on crystal growth most of which is in paper laboratory notebooks. Some of the data has been gathered in electronic form, e.g., the Biological Macromolecular Crystallization Database (BMCD) which is a repository of successful experimental conditions for growing over 800 different macromolecules (Gilliland 1987). Crystallographers are in need of computational tools to gather and analyze past data to design new crystal growth trails. We are building the Crystallographer's Assistant (CA) to help crystallographers record and maintain experimental context in electronic form, offer suggestions on experimental conditions that are likely to be successful, and provide explanations for failed experiments. As an initial step in this project, we have applied RL, an inductive learning program, to the BMCD. In this paper we report initial experiments and findings in applying RL to the BMCD. From the point of view of crystallography, we have discovered possibly significant new empirical relationships in crystal growth. From the point of view of machine learning, our work suggests refinements of existing methods for incorporating detailed domain knowledge into inductive analysis techniques.

Introduction

Starting with Sumner's crystallization of urease, crystals have been central to our understanding of biological macromolecules. Currently, 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.[1]

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, mostly in paper laboratory notebooks, 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.

In this paper we report on some initial experiments with a machine induction tool called RL (Clearwater & Provost 1990) with a large collection of crystallization trials – the Biological Macromolecule Crystallization Database (BMCD) assembled over several years by crystallographers (Gilliland 1987). We describe the unique features of the data domain that makes it a

*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] 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.

problem and the aspects of RL that makes it ideal as a data exploration tool.

An initial run using an off-the-shelf version of RL on the data showed the need for incorporating domain knowledge to guide the induction of rules. We discuss the use of data re-representation, attribute abstraction, and data subsetting, as a means of introducing general knowledge of the domain into the induction process. We show the following.

1. knowledge-based induction yields "better" rules than those generated by a statistical clustering analysis in this domain (SFR92). The rules discovered by RL on the BMCD data have revealed heretofore unknown correlations in the data that should be useful to protein crystallization.
2. as more domain knowledge is incorporated into the induction process, the "quality" of the derived rules improves. Unlike other domains where quality of rules is a single number (predictive accuracy on test set), it is a multi-dimensional object in this domain. We provide a analysis of how we can vary aspects of the quality of a rule set by varying some general parameters guiding rule generation.
3. general knowledge of crystallization can be used for attribute abstraction and data subsetting to yield rules at multiple scales that cover the data. To improve the quality of rules even further, we need to develop new methods for bringing the available knowledge on crystal growth, including qualitative theories of growth processes, to guide the generation and refinement of new empirical theories of crystallization from the experimental data.

From the point of view of crystallography, our contribution is the discovery of possibly significant new empirical relationships between experimental parameters. From the machine learning point of view, our work contributes to the refinement of methods for incorporating detailed domain knowledge into rule induction from data.

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.

Macromolecular crystallization

To be useful for crystallographic analysis, the molecule of interest must not merely be converted to the crystalline state, it must be converted into a single crystal with a high degree of internal order. These two parameters, size and order, are the critical measures of the success of a crystallization experiment because together they determine the quality of the X-ray diffraction pattern that can be obtained from the crystal. Unfortunately, neither is under the direct control of the experimenter. They can be manipulated by changing the "environmental" parameters of the experiment which include macromolecular concentration, pH, salt concentration, etc. To make matters worse, the precise relationship between the environmental parameters on the one hand and size and order on the other is generally different for each macromolecule. Thus each individual case has to be worked out empirically.

The basic scheme to induce crystallization is to slowly reduce the solubility of a sample solution of the macromolecule by one of several established methods. The solubility is determined by all the environmental parameters, one of which is usually the concentration of a "precipitating agent", such as polyethylene glycol. The "crystallization method", such as vapor diffusion, slowly raises the concentration of the precipitating agent (and almost everything else). If all the conditions are favorable, a point is reached where a crystal nucleates and grows.

The basic experiment is repeated with different parameters until either success is obtained or until the experimenter abandons the effort. Typically, many experiments (between 50 and 100) are started simultaneously and allowed to run for several weeks to several months. During this time the experimenter attends to other projects. Then, the results are evaluated and a new series begun (to run concurrently with the older ones). Thus, large volumes of data accumulate over long periods of time.

RL

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(BW93). We are using a machine learning program developed in our laboratory, named RL (CP90; PBC$^+$93), 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.

The RL program[2] is a direct descendant of the Meta-DENDRAL program (LBFL80), which established the soundness of viewing inductive learning as a knowledge-based problem solving activity that could be implemented in the heuristic search paradigm. It was first used to learn rules for predicting mass spectra of complex organic molecules, and has been gener-

[2] The exportable versions of RL are written in CommonLISP and C/C++ and run on any UNIX workstation with more than 4MB of memory.

alized and extended since then in several ways, which are mentioned below. Its method is primarily to search a space of possible rules, guided by prior knowledge and data in the training set. It learns a disjunctive set of weighted conjunctive rules[3]. RL has been used to explore databases and find rules in other problem areas including learning predictive rules for carcinogenic activity in chemicals (ALR+93), learning filtering criteria for high energy physics experiments (PB92a), and learning diagnostic rules for telephone line problems (PB92b).

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.

Rep	((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))
Interp	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. 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 statistic 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.

BMCD

The value of using experimental conditions from previous successful crystallizations as a guide to the design of new trials has been recognized in the experimental crystallography community. Gilliland (Gil87) has constructed a registry, called the Biological Macromolecule Crystallization Database which 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).

[3] i.e., associations of the form: if ($A1\&A2\&A3$) or if ($A3\&A4\&A5$), then conclude B.

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. In addition, the commercial version offers limited tools for analysis or design, it is simply a record of successful experimental conditions.

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.

Samuzdi, Fivash and Rosenberg (SFR92) 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 detail 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. 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 based on purely statistical methods.

Experiments with the BMCD data

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*.

Originally we intended to use the data in the BMCD to produce an initial, albeit weak, knowledge-base of rules for predicting a set of experimental controls from the parameters that are givens. This weak theory would then be used to both directly plan and guide the selection, adaptation and merging of experiments to plan new experiments. Unfortunately, the data in the BMCD does not adequately support this approach. Besides the lack of data about failures, there are very few givens in the BMCD other than the source and type of macromolecule and its molecular weight. This makes building a predictive expert system from this data difficult since there is insufficient information.

These features of the data forced us to readjust our goals. Rather than trying to induce a complete and coherent rule set connecting *givens* to *controllables*, we relaxed our goals to be that of verifying/discovering interesting and useful relationships from the available data. Notice that from a machine learning standpoint the second goal is vastly different from the original goal. First, it relieves us from the need to restrict the antecedents of our rules to a limited set of attributes available to the crystallographer prior to starting an experiment (*givens*). We can now base our rules on all the attributes in the data including those those that become available at the end of a crystallization experiment (*observables*). Second, we do not have to worry about finding a complete and coherent rule set that covers all of the examples. In fact, since we are looking for interesting relationships, well defined rules which may only cover a small subset of the entries may be preferable (an approach supported by Samuzdi's results). These types of rules would do a better job of precisely defining the context where the relationship holds, providing greater insight into the theory behind the relationship. By revising our goals we were able to transform an inadequate database into a corpus of data potentially capable of producing a set of both statistically and technically interesting relationships.

An important issue is how the rules generated by RL running on the BMCD data are to be evaluated. We used a mix of objective and subjective criteria for judging the quality of a rule set. The compression factor (the number of rules to the number of data points covered) was one of many factors in our assessment. It could not serve as the only means of judging the rules because the data set was too sparse, and the attribute set not complete enough for generating rules with high coverage. An important component in our assessment of rule sets was the number of rules judged to be "interesting" by the crystallographers. Crystallographers tended to rate rules that strengthen known relationships (confirmation rules) or those that suggested new relationships or contradicted existing lore (discovered rules) as "interesting". Crystallographers also value relationships between given and controllable parameters. Relationships between controllable parameters and observables can be used to bias experiment design. In our results below, we report the number of "interesting" rules found for different RL runs on the BMCD data.

RL is a complex induction engine with many tunable "knobs" and we used them to produce rule sets with higher compression factors. 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. 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 Samuzdi'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.

In order to analyze the BMCD data, one of the initial tasks was to convert it into a form usable by RL. The BMCD data has a lot of missing data and is not assembled for use by non-experts. This required us to do lots of data conversion that could be automated, although some of the fields required significant re-engineering. This reengineering took on four forms: data re-representation, attribute abstraction, data labeling 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 were members of the *chemad* attribute in the BMCD. Since RL does not handle set-valued attributes, we initially converted each additive into a real-valued attribute (moles per liter of that additive in the solution). Predictably, we obtained very poor rules since the number of attributes were comparable to the number of data points in the BMCD and the database was sparse to begin with. We replaced the *chemad* attribute by seven attributed which classified the chemical additives into seven classes based on their functional role in precipitation (buffer, slat, reagent, precipitant, 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. Hierarchical organization of attribute values was used to encode information about space groups. This allows RL to exploit the relationships between symbolic values of the data (e.g., I222, $P2_12_12_1$, and $C222_1$ are all orthorhombic forms). Uniform encoding of these values allows RL to produce more broadly applicable rules. An interesting open question for constructive induction in machine learning is whether these re-organizations of attributes and their values can be acquired automatically.

Data labeling, i.e., the identification of positive and negative instances in the data set, was used to analyze the nature of the BMCD data. As mentioned earlier, the BMCD contains only positive instances (i.e. successful crystallization attempts) while RL, as it is a supervised learner, needs both positive and negative instances to be able to generate rules. We 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., 7.0Å). We initially used a threshold of 3.5Å 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 3.0Å and 2.5Å to strengthen the rules. Table 2 shows the results on rule quality with varying diffraction limits set for labeling data as positive and negative.

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 ignore the attributes which have a coverage for less than 40% of the entries in the database. Also, in order to obtain more specialized relationships from the data, we have conducted experiments which use only subsets of the BMCD entries (e.g., only DNA or protein crystallizations). The data in Table 2 pertains to protein crystallizations alone.

Results

We now report results of the experiments with data labeling described above. Table 2 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* → *givens*, *controllables* → *controllables*, *observables* → *observables*, *givens* → *controllables*, *givens* → *observables*, and *observables* → *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 3.5Åwhile only 21 are generated with a limit of 3Å. However, the rules associated with a limit of 3Åare 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.

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:

We summarize 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 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.

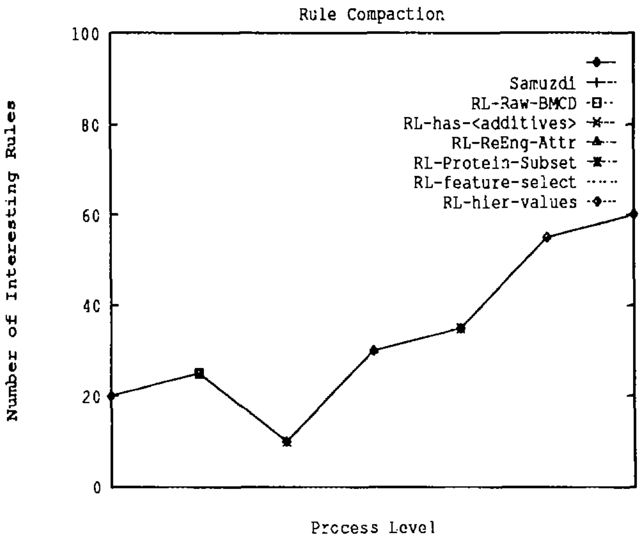

Figure 1:

A Sampler of Rules discovered by RL

We have processed the 1025 entries in the BMCD through repeated RL runs under several experimental settings described above. In this section we present "interesting" rules discovered by RL. A typical rule produced by RL has the following form:

> If the crystal habit had the value *plates* then the diffraction limit was under 3.5Å (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. The antecedents of the rules can have multiple conjuncts.

Included in the rules discovered by RL 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 instead of the raw RL representations.

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.
- 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 2.5Å) 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 3Å 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.

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 macromolecular concentrations (> 50 μg/ml) tended to have diffraction limits over 2.5$\AA$. 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[4]. 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. 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.

Future Work

Our initial work in applying RL to the BMCD suggests a number of areas requiring deeper study in order to extend our existing base of results. These fall under three broad categories: enriching the data representations, extending machine learning techniques to work with these representations, and incorporating additional forms of domain knowledge to guide induction.

1. Need for data on failed crystallization attempts, viz., data describing not just the successes, but also the experimental conditions of the failed experiments. This would allow more accurate and higher confidence rules to be inferred. In particular, it would give RL important additional information to judge when rules are being overgeneralized. The amount of data on failed experiments is much greater than that on successful ones. We are currently gathering data of this type in electronic form from Rosenberg's laboratory.
2. Need for more complete data on experimental context, 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 experimental conditions for the next batch of experiments. Treating sets of experiments as a unit will increase the reliability and credibility of the induced rules. We are capturing this experimental context in the data that is currently being collected.
3. Need for a multi-level representation of the experiments. The experiments themselves can be represented at multiple levels of detail. Hierarchical representations of experiments will allow RL to produce powerful, more broadly applicable rules in a more efficient manner.
4. Need for use of knowledge-based inductive learning techniques. One of the primary advantages of machine-learning over statistical analysis techniques

[4] A series of buffers that were developed by Dr. Good(GGW66)

is its ability to incorporate existing domain knowledge. The crystallography domain includes a significant corpus of knowledge about theoretical and experiential relationships that exist between experimental parameters that are manipulated in 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. The addition of such knowledge will enhance the performance of RL in terms of speed, accuracy and reliability.

5. 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 our project is to produce programs that would allow a typical crystallographer to use machine learning techniques without the help of a researcher in machine learning.

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

R. Ambrosino, Y. Lee, H.S. Rosenkranz, D.R. Mattison, F.J. Provost, and J. Gomez. 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.

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

S. Clearwater and F. Provost. Rl4: A tool for knowledge-based induction. In *Proceedings of the Second International IEEE Conference on Tools for Artificial Intelligence*, pages 24–30. IEEE CS. Press, 1990.

N.E. Good and et al. G.D. Winget. Hydrogen ion buffers for biologica research. *Biochemistry*, 5:467–477, 1966.

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

R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. *Applications of Artificial Intelligence for Chemical Inference: The DENDRAL Project.* McGraw-Hill, 1980.

F.J. Provost and B.G. Buchanan. Inductive policy. In *Proceedings of AAAI-92*, 1992.

F.J. Provost and B.G. Buchanan. 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.

F. Provost, B. Buchanan, S. Clearwater, Y. Lee, and B. Leng. Machine learning in the service of exploratory science and engineering: a case study of the rl induction program. Technical report, University of Pittsburgh, 1993. ISL-93-6.

C. Samuzdi, Fivash, and J. Rosenberg. Cluster analysis of the biological macromolecular crystallization database. *Journal of Crystal Growth*, 123:47–58, 1992.