Intelligent decision support for protein crystal growth

by I. Jurisica
P. Rogers
J. I. Glasgow
S. Fortier
J. R. Luft
J. R. Wolfley
M. A. Bianca
D. R. Weeks
G. T. DeTitta

Current structural genomics projects are likely to produce hundreds of proteins a year for structural analysis. The primary goal of our research is to speed up the process of crystal growth for proteins in order to enable the determination of protein structure using single crystal X-ray diffraction. We describe Max, a working prototype that includes a high-throughput crystallization and evaluation setup in the wet laboratory and an intelligent software system in the computer laboratory. A robotic setup for crystal growth is able to prepare and evaluate over 40 thousand crystallization experiments a day. Images of the crystallization outcomes captured with a digital camera are processed by an image-analysis component that uses the two-dimensional Fourier transform to perform automated classification of the experiment outcome. An information repository component, which stores the data obtained from crystallization experiments, was designed with an emphasis on correctness, completeness, and reproducibility. A case-based reasoning component provides support for the design of crystal growth experiments by retrieving previous similar cases, and then adapting these in order to create a solution for the problem at hand. While work on Max is still in progress, we report here on the implementation status of its components, discuss how our work relates to other research, and describe our plans for the future.

Proteins are involved in every biochemical process that maintains life in a living organism. One of the fundamental challenges of modern molecular biology is discovering the laws that control how proteins evolve their three-dimensional structure. Through an increased understanding of protein structure we can gain insight into the functions of these important molecules. Currently, the most powerful method for determining protein structure is single crystal X-ray diffraction.

A crystallography experiment begins with a crystal that ideally diffracts X-rays to high resolution, i.e., it produces a high-quality diffraction pattern that reveals the crystal's internal order. Crystals are regular, repeating arrays of atoms or molecules in three-dimensional space. The basic building block of a crystal is called a unit cell, the smallest unit of a lattice defined by three axes and the three angles between them. In order for a protein crystal to diffract at high resolution, it should not have large unit cell dimensions.

Determining protein structure is often limited by the difficulty of growing crystals suitable for diffraction. This is partially due to the large number of parameters affecting the crystallization outcome (e.g., purity of proteins, intrinsic physico-chemical, biochemical, biophysical, and biological parameters) and the unknown dependencies between the variation of these parameters and the propensity of a given macromolecule to crystallize. The primary goal of the research described in this paper is to develop a comprehensive repository of data from crystal growth experiments (both successful and unsuccessful) and apply this knowledge in an intelligent decision-support system for planning novel experiments.

0018-8670/01/$5.00 © 2001 IBM

Biomedical domains are characterized by substantial amounts of complex data, many unknowns, lack of complete theories, and rapid evolution. In such domains, reasoning is often based on experience rather than theory. Experts remember positive experiences for possible reuse of solutions; negative experiences are used to avoid potentially unsuccessful outcomes. Reasoning based on previous experiments (cases) provides a basis for a computational approach to problem solving known as *case-based reasoning* (CBR). A CBR system for solving crystal growth problems performs two major functions: (1) it retrieves "almost-right" prior crystallization conditions, which can then be modified to suit the new situation, and (2) it warns of potential errors or failures in proposed plans for crystal growth.

We start by first constructing a comprehensive repository of data from both successful and unsuccessful crystal growing experiments (our case base) using sophisticated robotic equipment that can carry out thousands of experiments a day. Moreover, the recording of the results from screening and optimization phases is automated. The stored cases will ultimately be used in conjunction with data-mining algorithms to derive general rules or principles related to crystal growth. Mining information on crystallization and using it intelligently is a challenge because of its multiple interdependent factors, the uncertainty of these dependencies, and the continuous evolution of our understanding of the data.

In summary, the objectives of the research described in this paper are:

1. Design and implement a comprehensive knowledge repository of crystal growth experiments (covering both positive and negative outcomes)
2. Using techniques from knowledge discovery and data mining applied to the repository of stored experiments, discover general principles for crystal growth and store this information in the knowledge repository
3. Design, implement, test, and evaluate a CBR-based intelligent decision-support system as an aid in the planning of crystal growth experiments

Specifically, we are developing an automated decision-support system for successful crystal growth that will help identify: the crystallization method of choice, the crystallizing agent of choice, the optimal temperature, the optimal pH, and the approximate concentrations of all solutes required in the crystal growth medium. CBR is an effective paradigm for such a system because: (1) it is similar to human expert problem solving and thus complements the user's decision-making processes; (2) it supports the evolving domain models and helps to increase domain understanding; and (3) it alleviates the problem of exceptions and over-generalizations.

The paper is organized as follows. In the next section we present an overview of protein crystallization and explain how crystal growth relates to current research in genomics and proteomics. Next, we introduce the architecture for Max, an automated decision-support system for crystal growth. Here we also describe its components: the information repository, the image-analysis component, the knowledge-discovery component, and the CBR component, and we report on the status of their implementation. We conclude with a discussion of related research and the potential impact of our work.

Protein crystallization

In this section we present the problem of crystal growth for proteins within the context of research in genomics and proteomics.

Genomics. The genome consists of threads of deoxyribonucleic acid (DNA). It contains instructions for making an organism, i.e., the blueprint for cellular structures and activities. The genome is organized into structures called chromosomes. Each cell in the human body has 23 pairs of chromosomes. A strand of DNA consists of repeating nucleotide units—a phosphate group, a sugar group, and a base: adenine (A), cytosine (C), guanine (G), or thymine (T). DNA is structured as a regular double-stranded helix, linked by hydrogen bonds between GC and TA bases. The human genome has about 3 billion base pairs.

A segment of a DNA molecule (a specific sequence of nucleotide bases) has a particular position on a specific chromosome. It carries information used for constructing proteins. Understanding how DNA performs this function requires an understanding of its structure and organization. The human genome contains about 100 000 genes. Genes have coding (exon) and noncoding (intron) sequences. Only about 5 percent of human genes are known to include the protein-coding sequences. The primary goal of the Human Genome Project[1] is to create detailed maps of each human chromosome. These maps have different levels of abstraction, dividing chromosomes into smaller fragments, and characterizing the fragments

and mapping them to their corresponding chromosome locations.

Once completed, the Human Genome Database will provide a unique scientific opportunity to researchers in biology, medicine, and computer science. However, a DNA sequence alone reveals little about protein function.

Proteomics. Proteomics involves the study of protein structure, function, and expression. Proteins are large, complex molecules composed of long chains of molecules, called amino acids. There are 20 common amino acids. A series of codons (triples of DNA bases) specify in which order amino acids are grouped to create specific proteins. Proteins provide the structural components of cells and enzymes for essential biochemical reactions.

Unlike the genome, which is identical in (almost) every cell of a particular organism, protein expression depends on a tissue, cell type, the stage of development, environment, and a disease state. Understanding protein function is necessary, as "Most disease processes and treatments are manifested at the protein level."[2] Thus, proteome analysis will significantly impact our understanding of the molecular composition and function of cells in both healthy and diseased organisms. Doctors may use this information to move from current medicine to individualized, molecular medicine. Based on a patient's genetic profile and the profile of the disease they will be able to custom-tailor treatment to an individual.

The structure of a protein is key to understanding its function. The three-dimensional structure arises from the folding of linear chains of amino acids into compact domains. In spite of considerable efforts to predict the structure of proteins directly from sequence information *in silico*, protein crystallography is currently at the forefront of methods for determining the three-dimensional conformation of a protein.

When a crystal is irradiated with X-rays, it scatters the radiation and produces a diffraction pattern. From this pattern (i.e., the collection of scattered rays) a three-dimensional picture of the atomic arrangement in the crystal can be obtained. Such structural information is crucial to our understanding of matter—if we know what a given molecule looks like, then perhaps we can understand and predict its properties.

High-throughput crystallization setup. The first step in determining the structure of a protein using X-ray crystallography is to grow a well-ordered crystal that is of sufficient quality to diffract X-rays strongly. A crystal is formed by numerous copies of a molecule becoming arranged in a tightly packed repeating motif. Well-ordered protein crystals are difficult to grow because proteins are large, irregularly shaped molecules that do not readily come together in a repeating pattern. Crystallization is a complex and tedious process; in some cases it may require months of trial and error to grow crystals suitable for X-ray diffraction analysis. This is because the formation of a crystal is critically dependent on a number of factors, including pH, temperature, protein concentration, the nature of the solvent and precipitant used, etc. Crystals form when molecules are slowly precipitated from solutions. This relates to a solubility diagram, which shows how increasing protein and crystallizing agent concentration results in moving from undersaturation through to the metastable zone, and from the nucleation zone to the precipitation zone. The saturated protein solution is in equilibrium with the crystallized protein. Small changes in any of the parameters may cause the protein to pack in different ways to produce different crystal forms. A set of crystal faces defines the crystal forms, which in turn define crystal morphology. However, this still does not reflect the overall shape of the crystal, which is given by the crystal habit. Recently, crystallization robots have been developed to automate and speed up the experimental process for crystallization.

In general, the number of experiments necessary to determine the optimal crystallization conditions is large, and often only a small amount of protein is available for crystallization. In our wet lab in Buffalo we now have the capacity to prepare and evaluate the results of over 40 thousand crystallization experiments a day. The experiments are automated using robots outfitted with syringes to dispense the cocktails (solutions or reacting agents) and protein, and a digital camera to record images (digital photographs) of the crystallization outcomes. The robotic setup enables us to maximize the number of initial experiments carried out for each protein (1536 conditions compared to the standard 48)[3] and minimize the amount of protein used. It also provides a controlled environment that promotes reproducibility of experiments and provides both successful and failed experimental results.

In the computer lab the recorded images from lab experiments are analyzed automatically to determine

the outcomes of the crystallization experiments. We are developing a standard vocabulary of outcomes to describe the results: clear drop, amorphous precipitate, phase separation, microcrystals, crystals, and uncertain outcome. It should be noted that for the purpose of high throughput screening, only conditions that result in crystals are worth keeping. However, for the purpose of case-based reasoning, all the conditions are used to form the *precipitation index*, our measure of similarity among proteins. A precipitation index is a vector in a space of 1536 dimensions representing crystallization outcomes for one protein and 1536 different crystallizing agents. Its binary representation has 0 for "clear drop" and 1 for "any precipitate" (we also distinguish unknown). A more refined representation further classifies precipitates into amorphous precipitate, phase separation, microcrystals, and crystal. Since precipitation is not an equilibrium, we evaluate each experiment five times over a few days. These outcomes, recorded as a function of time, are the cornerstone of our crystallization database. The database contains additional information, such as initial input data for the protein and the methodology (plan) used in carrying out the experiment. In the following section we discuss how a case-based reasoning algorithm can be used to identify patterns of similar properties and crystallization outcomes relating two or more proteins in the database.

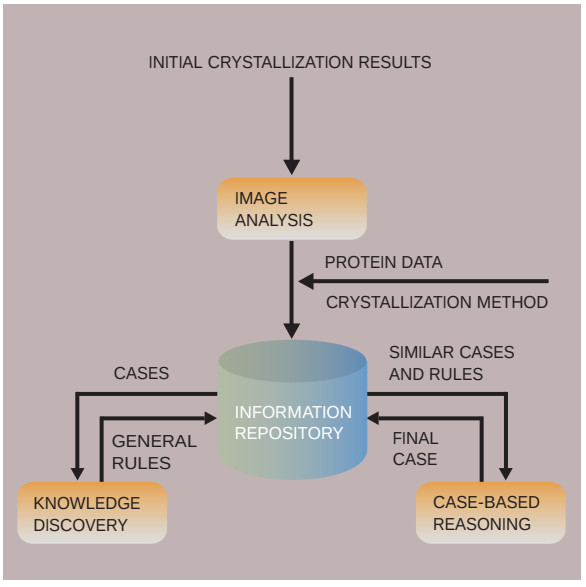
Max—Design and preliminary results

> *As the acquisition and availability of scientific data continue to escalate, the demand for improved bioinformatics, computer modeling, analytical tools, and remote access to research resources will necessarily increase as well.*
>
> —Recent statement by the National Center for Research Resources, USA

The information explosion in biomedical domains requires systematic knowledge management, i.e., support for acquisition, representation, organization, usage, and evolution of knowledge in its many forms. In this paper we focus on the issues of decision support and knowledge discovery for the protein crystallization domain.

Max is an intelligent decision-support system that is being developed to assist expert crystallographers in the planning of novel crystal growth experiments. As illustrated in Figure 1, Max comprises three computational components: an image-analysis component to automatically evaluate the outcome of crystallization experiments, a case-based reasoning component for the design and analysis of experiments, and a knowledge-discovery component to help discover the underlying principles of crystal growth. The database component of Max is an information repository that includes some existing databases and newly created data based on experiential knowledge and general principles and rules of crystal growth. In the remainder of this section we describe each component in turn and report on the status of its implementation.

Figure 1 System architecture



Information repository. At the conceptual level, the information repository contains both data and knowledge. Data comprise existing databases—verified information from the Protein Data Bank[4] (PDB), the Biological Macromolecule Crystallization Database[5] (BMCD), and GenBank[6]—as well as specialized domain information about proteins, their structures, functions, chemicals, and reacting agents. Knowledge in the system's repository has two forms—experiential (cases consisting of information about individual past experiments with diverse crystallization outcomes) and general (principles or rules that are derived from the knowledge-discovery system or through knowledge acquisition).

The information repository is being created systematically with an emphasis on information quality (i.e.,

correctness, completeness, reproducibility), its current focus being on the development of a case base for storing experiential knowledge. Aside from its use as a component of Max, the information repository will be made available to other crystallographers. In the remainder of this section we describe the data included in a case, along with related data management issues.

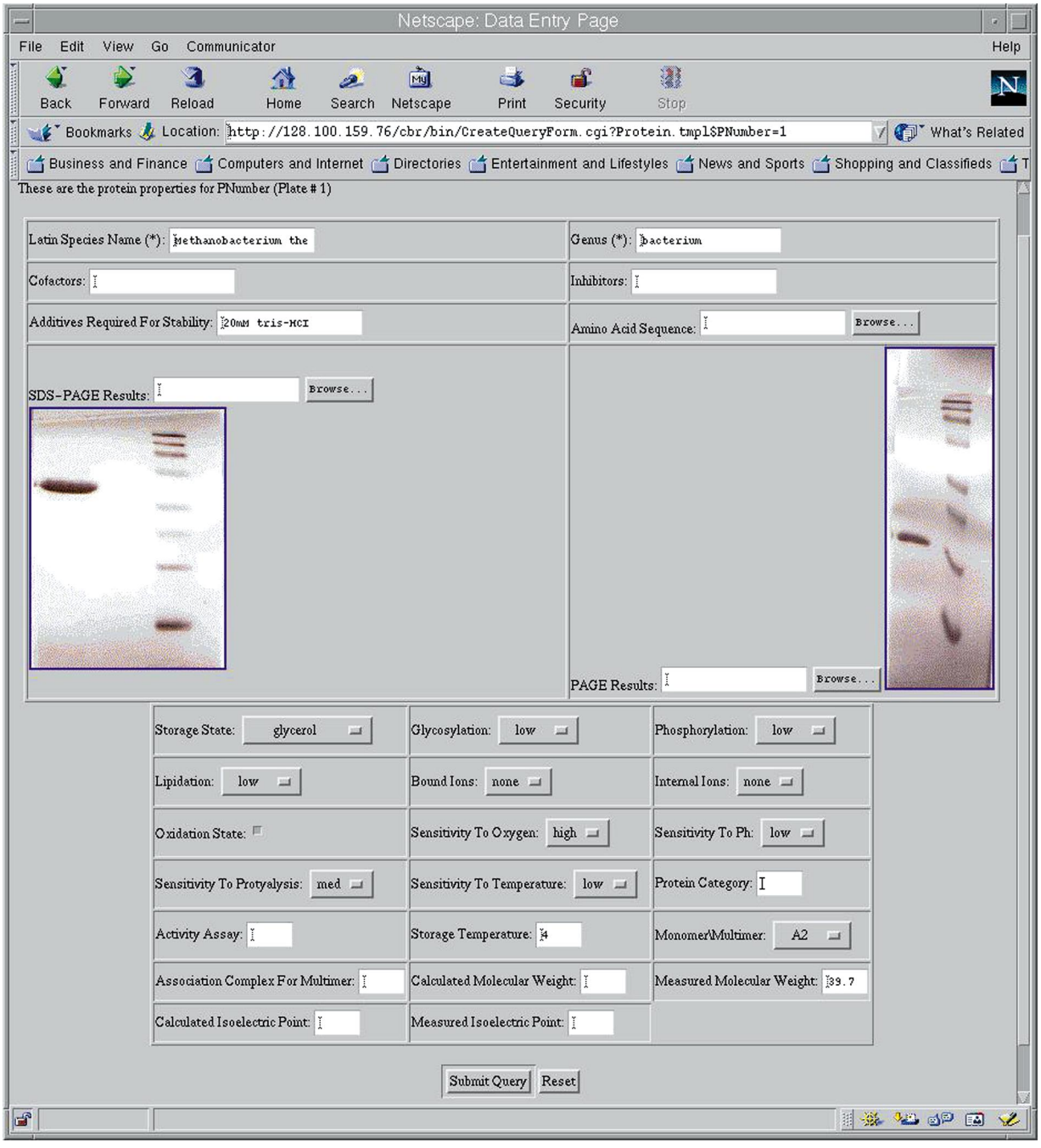
In general, a case consists of data related to a specific task and its outcome. In Max, a crystallization case for a given protein captures the problem-solving process of a crystal growth experiment, which includes biochemical properties (the problem), crystallization conditions (the solution), and crystallization experiment outcomes (the feedback), protein properties (including amino acid sequence, species), ancillary biophysical information such as gel scans, the results of the crystallization in the form of image and extracted features, a preliminary classification of the experiment outcome in the form of a precipitation index, the crystallization method, the crystallizing agent, the optimal temperature and pH, approximate concentrations of all solutes required in the crystal growth medium, and the experimental outcomes of optimization trials (both successes and failures). A crystallizing agent is always a precipitating agent, but the reverse does not necessarily hold. As described below, our use of high throughput screening is to get protein precipitation reactions in 1536 different conditions, and we equally value precipitates and crystals. It should be noted, however, that if a particular condition always produces a precipitate or a clear drop result, we will eliminate it since it does not help us to differentiate among different proteins.

Information about a particular protein will be entered into our repository only if it was crystallized at least once. In addition to the above-mentioned properties, a case also stores information about the diffraction experiment. Although each protein in the repository had to be successfully crystallized at least once, this does not mean that the crystal diffracts X-rays well. There might be multiple crystals available, of which only one will be suitable for diffraction. Storing this information in a case will enable the reasoning system to use it during adaptation to prioritize successful (from the X-ray diffraction experiment viewpoint) recipes. The case-based reasoning paradigm can help only if we can draw on successful and failed crystallization experiments. Figure 2 depicts a Web-based data entry facility for case authoring. The cases are stored hierarchically in a DATABASE 2* (DB2*) database management system running on an IBM RISC System/6000* (RS/6000*) SP2*. As will be discussed in a later section on knowledge discovery, the case base is organized according to attribute categories in order to promote efficient retrieval during similarity matching.

Initially, in order to populate the repository, the wet lab is conducting thousands of crystallization experiments, which are automatically evaluated and all relevant properties captured. It takes about ten minutes to robotically prepare the plate and about 20 minutes to digitally capture experimental outcomes. Cocktail and protein properties are captured before the experiment via a Web interface and stored in the database. Additional information will be added to the repository, once the experiment is completed.

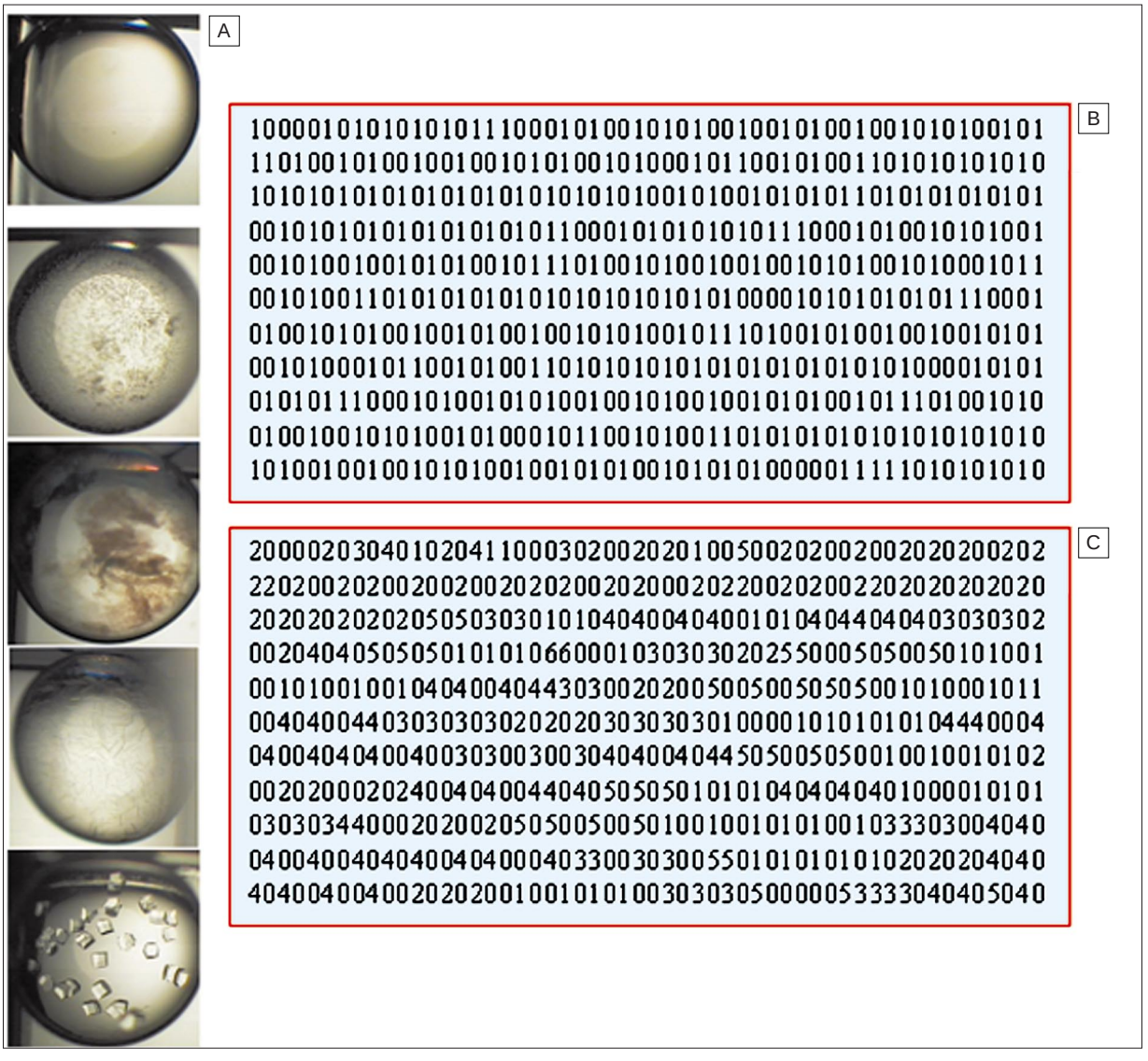
Image analysis. This subsection describes an image-analysis system used to automatically classify crystallization experiment outcomes. The motivation to build such a system is twofold: (1) there is no general approach to quantitatively and objectively evaluate reaction outcomes under the microscope, and (2) there is a need to eliminate human intervention in order to cope with the high-throughput of the robotic setup. The major weakness of existing scoring methods is the tendency to confuse categories of precipitates.[7] As previously stated, we store crystallization outcomes as images, analyze them using computer vision techniques, automatically recognize the possible crystallization outcomes, and extract important image features for further analysis.[8,9] It is important to note that such a process produces objective results, which have the potential to be incorporated in the data-mining process. Recall that for each protein, experiments are carried out using 1536 different cocktails. This forms a "signature" of a protein known as the *precipitation index* because it defines the precipitation properties of a protein. The precipitation index allows us to measure similarity between proteins; our hypothesis is that the closer the indices are, the more likely it is that the two proteins will have similar crystallization plans. Experiment outcomes, recorded as functions of time, are the cornerstone of the crystallization case base and are used to retrieve "similar cases" during CBR (the process of similarity-based retrieval is explained in the next subsection). The task for image analysis is to determine the outcome of each of the 1536 different experiments. The result of a precipitation experiment is a robotically captured set of images that can be analyzed to determine the outcome of the crystallization process. We are developing a standard vo-

Figure 2 Web-based data entry of protein information



cabulary for the crystallization process, where each image is classified as one of: clear drop, amorphous precipitate, phase separation, microcrystals, crystals, and uncertain outcome. Figure 3 illustrates several crystallization experiment outcomes, a partial example of a binary encoded precipitation index, and a

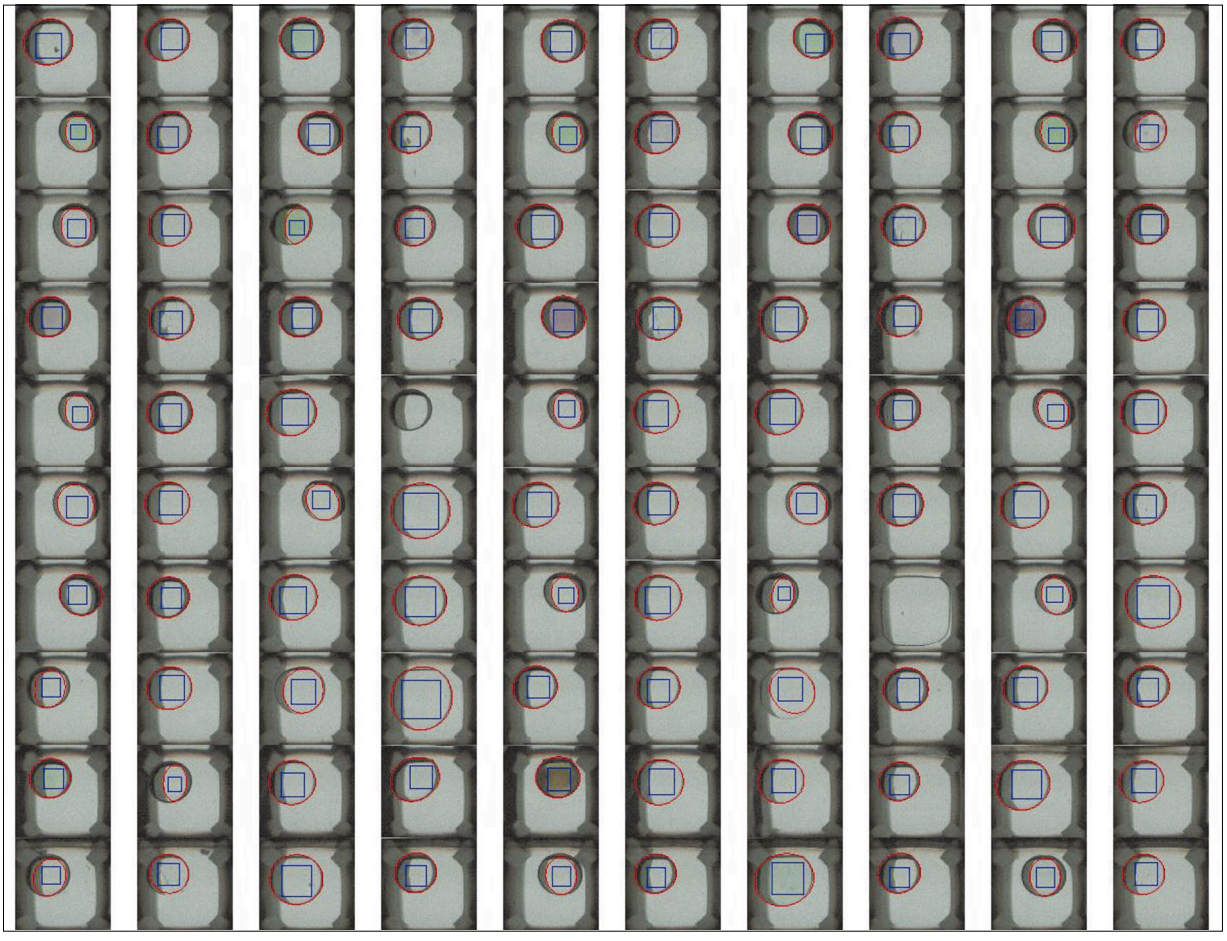
Figure 3 Precipitation index. (A) illustrates 5 of the 1536 images for a given protein. (B) gives a binary classification consisting of 1s (something happened) and 0s (nothing happened) for the set of images. These can further be refined to the more detailed outcomes that are encoded in (C). Currently, we distinguish clear (0), crystal (1), precipitate (2–4), and unknown (5). Once the classification accuracy is improved, we will differentiate between amorphous precipitate (2), phase separation (3), and microcrystals (4).



corresponding precipitation index that differentiates among precipitates. All images are processed on the Toronto University's IBM RS/6000 SP2 using MATLAB**.[10] Currently, we process each image in about 0.5 seconds, which matches the rate of the high throughput production at our Buffalo lab. The average error rate of the drop boundary recognition is 0.4 percent and experiment outcome classification 85 percent.

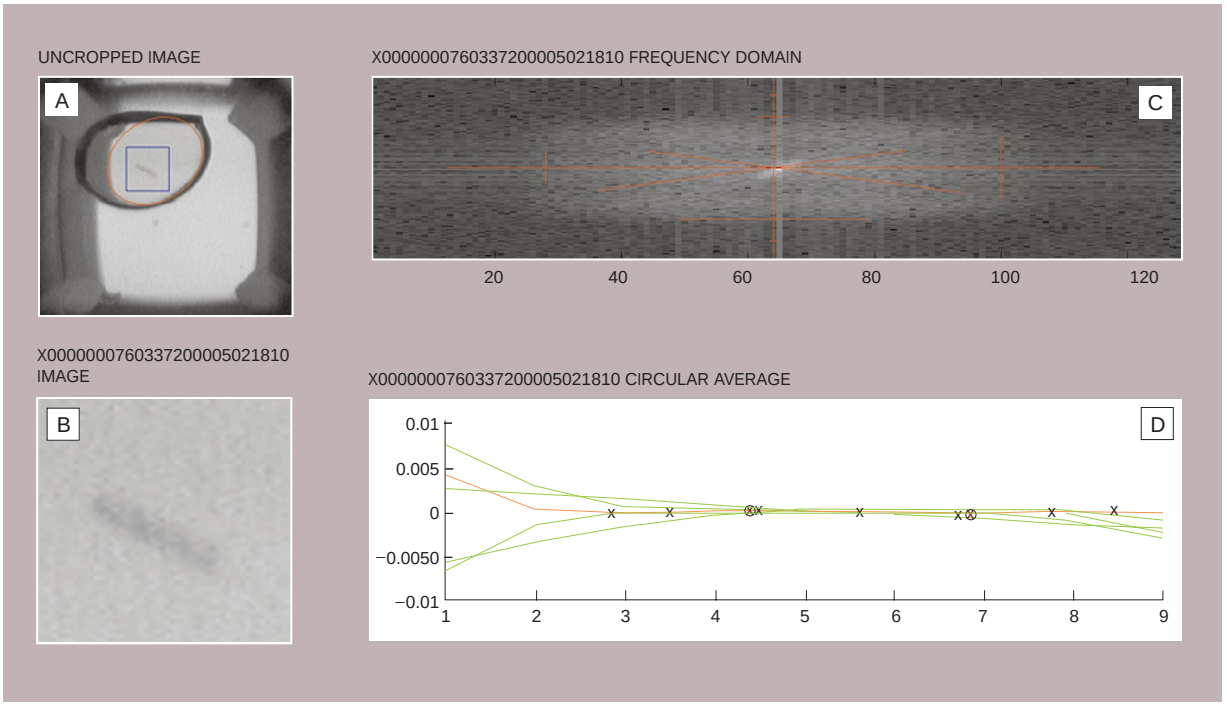
An image preprocessing step attempts to standardize images with regard to lighting, size, and orientation. Image processing involves three primary steps:

Figure 4 Droplet identification



1. *Droplet recognition:* Figure 4 illustrates the automated determination of a drop within each well, and the identification of the largest square inside the drop. First, the best conic is fit to the boundary of the droplet and then the droplet's largest square area is identified for further analysis, as shown in detail in Figure 5A and 5B. Because the images are not trivial and the drop can have many different forms, there are multiple "feasible" drop boundaries. The image-analysis subsystem generates an ordered set of viable alternatives for recognizing the drop boundary. The most probable boundary is the conic with the most likely shape, size, and position.
2. *Analysis and feature extraction:* The second step in image analysis is to determine the characteristic properties of the image. This is needed to support automatic classification of the experiment outcomes and to enable the comparison of results from crystal growth experiments. We use the two-dimensional Fourier transform to perform image analysis. The frequency domain of the Fourier transform is presented in Figure 5C, whereas Figure 5D depicts an analysis of the spectrum derivatives and circular averages. These analyses provide important feature information for the image. Figure 6 illustrates a portion of a case representation comprising an image and 35 derived image features for the given experimental outcome (e.g., vertical height, horizontal width, left/right spur height, left/right spur brightness ratios, the number of elements in a quadtree decomposition with different thresholds, the location of the first change in curvature of the circular average of the frequency spectrum).

Figure 5 Image analysis by Fourier transformation. (A) shows the recognized drop. (B) shows the drop's largest square. (C) presents the spectrum of Fourier analysis. (D) depicts analysis of the spectrum.



3. *Classification:* The third step in processing an image is to classify the outcome of the crystallization experiment. Initially we determine whether something or nothing happened and, if something happened, then attempt to refine the classification into one of the possible outcomes. Figure 7 illustrates three robotically captured images of experiments and the resulting spectral analysis for these. Since the spectrum characteristics correlate with the experiment outcome, the three images can be classified as clear drop, amorphous precipitate, and microcrystals, respectively. According to Carter, fluffy or filamentous precipitates have little likelihood of being crystalline, but uniform, granular, and/or particulate precipitates often are microcrystalline.[11] Our preliminary results show a possibility to automatically distinguish between these precipitates; we are currently working on process optimization to increase the accuracy.

We are in the process of implementing a distributed storage management system to help us cope with the increasing volume of image data and to support archiving of important information (we already have 150 GB of compressed images containing crystallization experiment outcomes). The system comprises an LTO (Linear Tape Open) tape library attached to the IBM SP2, the IBM ADSTAR* system, and the IBM DB2 EEE database. Images are transferred from Buffalo to Toronto via a fast Internet2 connection. Image features are extracted from the high-resolution images and stored with other experiment information in the database (although original images are in TIFF [tagged image file format] format, we have experimentally established appropriate JPEG [Joint Photographic Experts Group] compression that does not affect the analysis). Since the image-feature extraction algorithm is being improved to increase classification accuracy and the imaging settings also change over time, we need versioning of images and corresponding MATLAB code. As described earlier, case retrieval does not use the raw image, so the images can be stored in the tape library and accessed for final validation of retrieval results and for batch processing for feature extraction.

Case-based reasoning. Biological domains often require multimodal representations that support both textual and pictorial data. Although diverse tools are necessary, we apply CBR as the core technology for our project because it uses experiential knowledge as a guide to problem solving. CBR generally involves adapting old solutions to meet new demands, or using old cases to explain or critique new solutions.[12]

The process of crystal growth can be considered as a planning task, where a single experiment corresponds to a simple plan and a series of experiments for a given protein corresponds to a more complex plan. Our approach builds upon a previously developed computational framework for CBR called TA3.[13] This system employs a variable context, a similarity-based retrieval algorithm, and a flexible representation language. Cases, corresponding to individual experiences, are stored in TA3 as a collection of attribute-value pairs; attributes are grouped into one or more categories to bring additional structure to a case representation. This reduces the impact of irrelevant attributes on system performance by selectively using individual categories during case retrieval. Figure 6 depicts a partial representation of a case for a given protein experiment. As illustrated, the case stores values for properties such as the molecular weight of the protein and the temperature used in the final plan for crystallization. The precipitation index is depicted as an array, where check marks denote that some activity occurred for the corresponding crystallization experiment (see Figure 8).

Statistical analysis plays a major role in identifying the significance of individual descriptors (cocktails in precipitation index, protein properties, and features extracted from experiment images), and in determining how to automatically relax queries during iterative retrieval; correlation between attributes is used to group them into categories, and value histograms are used to guide the query relaxation process (generalization). We use confidence measures during the adaptation process, to guide the generation of a crystallization plan from previous successful and failed crystallization experiments for similar proteins. Statistics will also be used once we start applying knowledge-discovery techniques to the case base to measure support of, and confidence in, the discovered patterns.

As illustrated in Figure 9 there are several processes involved in CBR. Case retrieval involves partial pattern matching of the crystal growth problem to cases

Figure 6 Example of a crystallization case. Note that only experiment outcome and extracted image features are shown here.

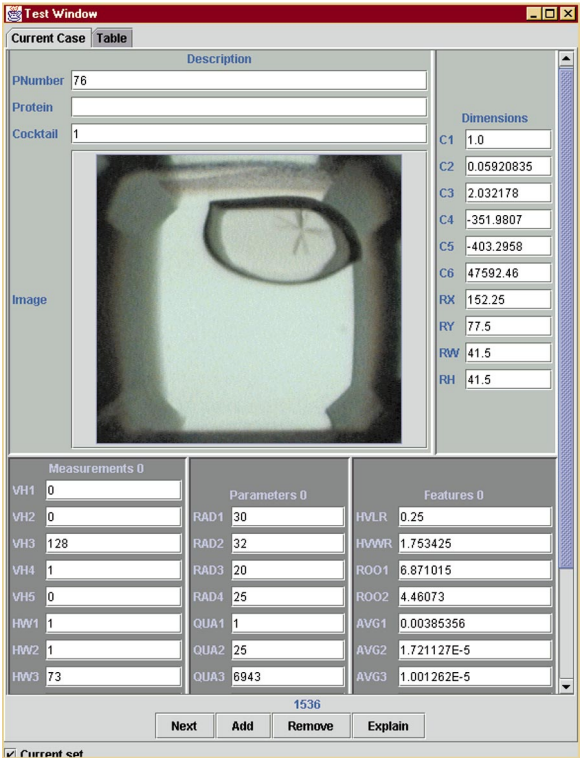


Figure 7 Automatic classification of the experiment outcome

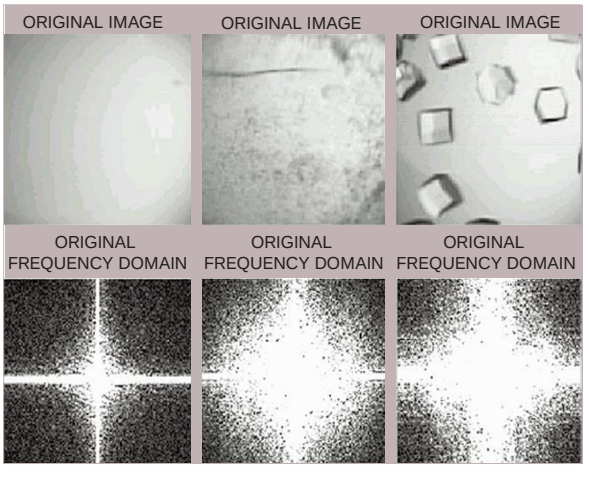


Figure 8 Basic protein properties and part of the binary representation of the precipitation index

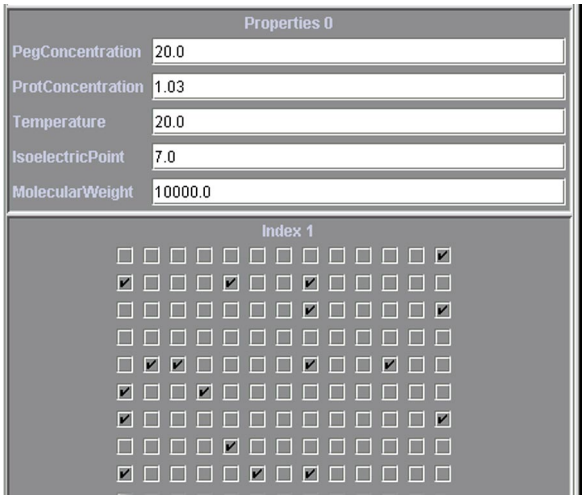
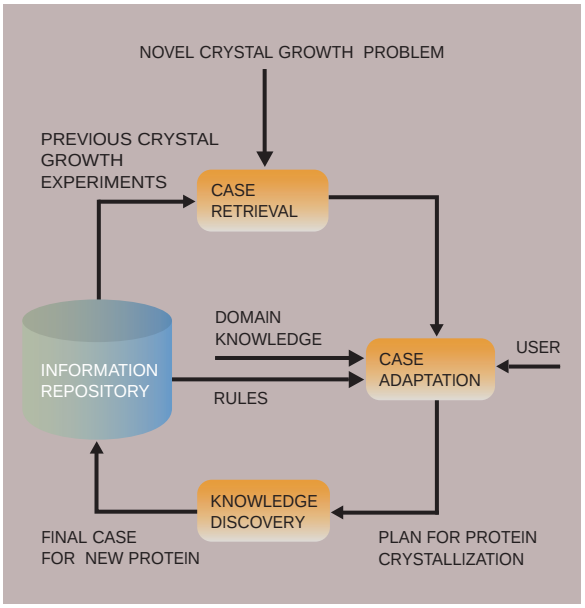


Figure 9 Case-based reasoning process for Max



in the case base. A similarity function is used to determine which cases are most relevant to the given problem. Initially we are using the precipitation indices to define a quantitative similarity function. We postulate that past experience can lead us to the identification of initial conditions favorable to crystallization. Moreover, it is hypothesized that solubility experiments can provide a quantitative measure of similarity among proteins. Since we use high throughput screening to get a precipitation index, we define solubility experiment as the amount of compound dissolved in a solution in the presence of precipitate. Each plate is evaluated five times to see how the precipitates change over time. Assume that two proteins react similarly when tested against a large set of precipitating agents. Then the crystallization strategies successfully employed for the one may be profitably applied to the other. Thus, we must identify a suitable set of precipitating agents to sort the outcomes of reactions for a relatively large group of proteins, all of which have been crystallized before. New crystallization challenges are then approached by the execution and analysis of a set of precipitation reactions, followed by an automated identification of similar proteins and analysis of the conditions used to crystallize them.

Once similar cases have been retrieved, the next step in CBR is adaptation. This is the process of modifying previous solutions to address the new problem. The most relevant retrieved cases, along with domain knowledge, are incorporated to determine the most appropriate parameters for our new set of experiments. At this stage the system acts, first, as an advisor to the crystallographer to suggest possible parameters for further experimentation and, second, as an evaluator of potential experiments that the user might propose. The adaptation module will be enhanced over time as general rules/principles extracted from the growing case base are used in the adaptation algorithm.

The final step of our crystal growth problem is to evaluate the outcome of the crystal growth experiment and to author an appropriate case (containing information about the protein, crystallization plan, outcome, etc.) in the information repository. Further details on the case retrieval and case adaptation processes for CBR follow.

Case retrieval. Case retrieval involves a pattern-matching algorithm in the similarity-based search of the case base. The retrieval is based on a modified k-nearest neighbor algorithm,[14] where: (1) attributes are grouped into categories of various importance levels to help control the matching process and reduce the negative impact peripheral attributes have on performance; (2) an explicit context is used during similarity assessment to ensure that only relevant cases are retrieved, that the process is visible to the

user, and that the user has the ability to intervene by introducing bias; and (3) incremental context transformations are applied during query relaxation to speed up query processing.

In the Max system, the case retrieval function is used to locate crystallization experiments that have similar precipitation indexes. Thus, the precipitation indices are used to define a quantitative similarity function. The retrieval process has two stages. In the first stage, only a binary classification of crystallization outcomes is used (i.e., nothing happened, something happened). In the second stage, a more detailed classification of the result is used to partially order retrieved experiments based on their relevance. In both steps the system evaluates the distance between the precipitation reaction index for a novel protein with those in the case base. Only cases with minimum Hamming distance are considered in the second step. Other attributes, such as protein sequence, weight, etc., are considered only as auxiliary information during the retrieval process.

We incorporate a context-based retrieval method,[13] implemented in the TA3 system, to allow the user a flexible interface for restricting or relaxing the context in order to retrieve fewer or more cases as necessary. The context is used to specify what attributes are important for retrieval and what ranges of values for these attributes are allowable for determining similarity between the input case and stored cases. The retrieval process is geared to interactive use and applies an efficient incremental algorithm.[15]

An explicitly defined context controls the closeness of retrieved cases. If too many or too few relevant cases are retrieved using the initial context, then the system automatically modifies the context unless the user does it manually. Modifying the context controls the quality and the quantity of retrieved cases. Depending on the change in the context, the system may return an approximate solution quickly or it may take longer to produce a more accurate solution. An approximate answer can be iteratively improved, so that the range between an approximate and an accurate answer is a continuum, an important feature for bounded-resource computation.[16,17]

TA3, the underlying CBR framework for Max, uses two context transformations to support iterative retrieval and browsing: *relaxation*, which retrieves more cases, and *restriction*, which retrieves fewer cases. Context relaxation is applied either by *reduction* or by *generalization*. Reduction removes an attribute-value pair from a context, either permanently or dynamically; given x-of-n matching, the number of attributes required to match is reduced from x to y, where $0 < y < x \leq n$. Generalization relaxes the context by enlarging the set of allowable values for an attribute. Contexts can also be iteratively restricted to retrieve successively fewer cases. Context restriction is applied in two possible ways: *expansion*, i.e., strengthening constraints by enlarging the number of attributes required to match, and *specialization*, i.e., strengthening constraints by removing values from a constraint set for an attribute. The implementation of an interactive context modification is further described in Reference 15.

Retrieved cases are presented to the user, at which time the user can modify the selection criteria dynamically and thus alter the set of cases to be retrieved next. The retrieval process is interactive and iterative. The retrieval function used in Max is flexible, effective, efficient, and scalable.[13,15] The higher the precision and recall of case retrieval the easier and more accurate the case adaptation process.

Case adaptation. Once relevant cases have been retrieved, the next step in CBR is adaptation. This is the process of modifying a previous solution to address the new problem. Adaptation in CBR manipulates the existing solution to better fit the target case. Our hypothesis is that, given a quantitative measure of similarity between proteins (in this case the precipitation index), recipes successfully employed for one protein will be useful as starting points for crystallization experiments for similar proteins.

In Max, the cases retrieved in a similarity search, together with the domain knowledge stored in the information repository, are used to determine the appropriate parameters for the new crystallization experiment. The solution is a recipe for crystallization, i.e., crystal growth method, temperature and pH ranges, concentration of protein, and crystallization agent. The system acts as advisor and evaluator by finding and using relevant successful and unsuccessful experiments in the case base. Max advises the crystallographer by suggesting possible parameters for further experimentation. It evaluates potential experiments that the user might propose, and warns of potential failure based on previous experience.

Adaptation is guided by domain knowledge, i.e., adaptation rules, concept hierarchies, cases stored in the Max information repository, or information pro-

vided by the user. If the information is provided by the user, Max remembers it for possible later reuse. Once a new set of experiments for the target protein has been executed, a new case, which reflects this new experience, is added to the case base.

Knowledge discovery. Several terms have been put forth[18] to describe the process of finding useful patterns in data. These include data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term "knowledge discovery" was introduced at the first Knowledge Discovery in Databases workshop in 1989. Similar to data mining, knowledge discovery emphasizes the end product of the discovery process, which is knowledge. It is not sufficient that the pattern is novel, it must also be in a form that the human users will be able to understand and use. Knowledge discovery has been defined as "a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."[19] The process of knowledge discovery employs methods from statistical data analysis and machine learning.

One of our research objectives is to search a mature crystal growth case base looking for interesting and unanticipated relationships. Using data visualization tools and formal knowledge-discovery algorithms for numeric and conceptual cluster analysis, we hope to uncover new trends in the data from crystallization experiments that can be exploited as we face new crystallization challenges. Toward this goal we will pursue the integration of a loosely coupled federation of biological databases. In addition, we will explore relationships between features of experiments and characteristics and/or properties of the protein, e.g., sequence and structure.

The knowledge discovery in Max has two distinct features. First, it is based on a multimodal representation, i.e., it uses not only numeric attribute values, but also symbolic values and still images. Second, its tools support multiple tasks, including optimization, domain knowledge evolution, and adaptation support. Optimization involves locating descriptors relevant to a given context and task. These descriptors are then used to determine the most appropriate attributes for case retrieval, and to determine the appropriate constraint values (context) for an attribute. This helps determine the features of a crystal growth experiment that are predictive of outcome (positive or negative). Domain knowledge evolution involves analyzing the case base to determine the best attributes/descriptors for CBR. Extending the case representation with additional descriptors improves case discrimination during prediction and classification. Removing redundant cases and descriptors also improves performance since only relevant descriptors are used during similarity-based retrieval. Creating hierarchies of descriptors enhances knowledge organization and thus improves system performance and domain comprehensibility. Analyzing created clusters, hierarchies, and associations may lead to identifying underlying principles for the problem of crystal growth and to the discovery of nonintuitive relationships among features in a crystal growth experiment. All information sources pertinent to a crystallization experiment will be considered, including cases, images, and agent and protein databases. This may lead to finding correlations among precipitation cocktails based on the protein chemical and physical properties, e.g., it may explain why a certain group of proteins reacts similarly with respect to a particular set of cocktails and how these cocktails relate. The discovered principles will be used to support case adaptation in the CBR system.

We are currently working on providing integrated access support to several biological databases.

Discussion

The design and implementation work on Max is still in progress. A Web-based interface and a relational schema to store data from crystallization experiments have been implemented. Currently, we are working on integrated access support to other biological databases, such as PDB,[4] SCOP,[20] and SPINE.[21]

The results of the crystallization experiments are examined visually using a computer-controlled XY table and a videomicroscope that feeds a framegrabber. The XY stand can accommodate 28 plates of experiments, allowing us to photograph $28 * 1536 = 43008$ experiments in about 9 hours. Each photograph is saved as a JPG image (320×320 pixels in RGB, or red, green, blue). Photographs are taken on a regular basis: immediately following setup, one day later, two days later, one week later, and two weeks later. Currently, we are working on extending the

imaging setup. We obtained a new, higher resolution camera (2 Mega pixels), and we are experimenting with taking images both from top and bottom of the well. Our preliminary results show that we will be able to improve classification accuracy significantly, due to better differentiation of image features.

We have implemented TA3 in Java** 2, with both memory and JDBC** (Java Database Connectivity) drivers. Cases can be stored in a hierarchical manner to support more efficient storage (because one protein may be part of multiple crystallization experiments), improved case retrieval performance, and knowledge discovery through exploiting meaningful structure of case base. We are working on improving its performance and extending its knowledge-discovery capabilities. Currently, knowledge discovery supports only case similarity explanation and TA3 optimization by case schema refinement and domain knowledge analysis.

Current approaches to the crystallization of macromolecules are primarily empirical. Because of its unpredictability and high irreproducibility, crystal growth has been considered by some to be an art rather than a science.[7] Even so, experimental protocols for crystal growth that are effective in many settings have evolved. For example, Jancarik and Kim proposed a set of 48 agents that are often used during crystallization.[3] In spite of the progress made, there remains a need for systematic studies to improve our understanding of the crystallization process and to support the design of successful new experiments. This need is compounded by the promise of genomics projects to produce hundreds of proteins a year for structural analysis.

An additional problem in crystal growth has been a historically nonsystematic approach to knowledge acquisition: "the history of experiments is not well known, because crystal growers do not monitor parameters."[7] BMCD[5] stores data from published crystallization papers, including information about the macromolecule itself, the crystallization methods used, and the crystal data.

Others have attempted to apply machine learning techniques to the BMCD database. These efforts include an approach that uses cluster analysis,[22,23] an inductive learning method,[24] and correlation analysis combined with Bayesian technique[25] to extract knowledge from this existing database of crystallization experiments. These studies were limited because negative results are not reported in the database and because many crystallization experiments are not reproducible due to an incomplete method description, missing details, or erroneous data. Consequently, BMCD is not currently being used in a strongly predictive fashion. The information repository of crystal growth experiments we are developing addresses these shortcomings.

Decision-support systems have previously been considered in a variety of applications in molecular biology: e.g., to help identify protein secondary structures,[26] to assist in locating molecular motifs,[27] to find similarity between protein structures,[28] and to help during the initial stage of drug discovery.[29] Some of the fundamental problems that have to be addressed when applying artificial intelligence techniques to the molecular domain are: how to effectively represent information,[30] how to access it effectively and efficiently,[29] how to analyze it,[31] and how to reason with it during decision making. Given the uncertainties present—the diversity of representational formalisms used, the complexity and amount of information present, and the evolution of knowledge—it is necessary for an intelligent information system to be flexible and scalable.

Future research in the development of Max includes implementing a parallel version of TA3. We also plan to apply TEIRESIAS[28] to the knowledge-discovery process. TEIRESIAS is a system that has successfully been applied to other application domains in molecular biology (including homology search, multiple sequence alignment, and the discovery of tandem repeats in DNA sequences). TEIRESIAS can also be applied to the more general problem of association discovery in data sets that come from a variety of scientific domains. Carrying out association discovery is the first step toward discovering causal relationships and predicates that are expected to be particularly useful in the context of our work on protein crystallization.

The work described in this paper offers several research challenges. First, it involves developing and applying advanced data-mining and knowledge-discovery techniques to a complex scientific domain. It also requires the integration of image data with CBR, along with the application of sophisticated knowledge management tools. In the past, researchers in artificial intelligence have often been criticized for restricting their applications to "toy problems"; crystal growth is a complex, real-world domain where an intelligent decision-support system has dramatic impact. Improving and accelerating the process of

protein crystal growth will aid in the expansion of the repository of known structures. The significance of this is far-reaching: increased knowledge of protein structure is critical to medicine, drug design, and enzyme studies, and to a more complete understanding of principles in molecular biology. Due to the need to accelerate crystal structure determination to take advantage of the wealth of information arising from genomics projects, it is particularly crucial at this time to improve the process of crystal growth.

Acknowledgments

The computing part of this research is supported in part by the Natural Sciences and Engineering Research Council of Canada, Communications and Information Technology Ontario, and IBM Canada; the wet lab is supported in part by the John R. Oishei Foundation and NASA Grant NAG8-1152. Both the computational lab and the wet lab are supported in part by the NIH grant to the Northeastern Structural Genomics Consortium (http://www.nesg.org).

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of The Mathworks, Inc., or Sun Microsystems, Inc.

Cited references

1. D. R. Bentley, "The Human Genome Project—An Overview," *Medical Research Review* **20**, No. 3, 189–196 (2000).
2. J. H. Wang and R. M. Hewick, "Proteomics in Drug Discovery," *Drug Design Today* **4**, No. 3, 129–133 (1999).
3. J. Jancarik and S. H. Kim, "Sparse Matrix Sampling: A Screening Method for Crystallization of Proteins," *Journal of Applied Crystallography* **24**, No. 4, 409–411 (1991).
4. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research* **28,** 235–242 (2000).
5. G. L. Gilliland, M. Tung, D. M. Blakeslee, and J. Ladner, "The Biological Macromolecule Crystallization Database, Version 3.0: New Features, Data, and the NASA Archive for Protein Crystal Growth Data," *Acta Crystallographica* **D50**, 408–413 (1994).
6. D. A. Benson et al., "GenBank," *Nucleic Acids Research* **28**, No. 1, 15–18 (2000).
7. *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, A. Ducruix and R. Giege, Editors, Oxford University Press, New York (1992).
8. J. I. Glasgow and I. Jurisica, "Integration of Case-Based and Image-Based Reasoning," D. W. Aha, Editor, *AAAI'98 Workshop on Case-Based Reasoning*, Madison, WI, AAAI Press, Menlo Park, CA (1998), pp. 67–74.
9. I. Jurisica and J. I. Glasgow, "Extending Case-Based Reasoning by Discovering and Using Image Features in IVF," *ACM Symposium on Applied Computing* (*SAC 2000*), Villa Olmo, Como, Italy, ACM, New York (2000), pp. 52–59.
10. See http://www.mathworks.com.
11. C. W. Carter, Jr., "Design of Crystallization Experiments and Protocols," *Crystallization of Nucleic Acids and Proteins*, A. Ducruix and R. Giege, Editors, Oxford University Press, New York (1992), pp. 47–71.
12. J. L. Kolodner, *Case-Based Reasoning*, Morgan Kaufmann Publishers, San Mateo, CA (1993).
13. I. Jurisica and J. I. Glasgow, "Improving Performance of Case-Based Classification Using Context-Based Relevance," *International Journal of Artificial Intelligence Tools*, *Special Issue of ICTAI'96 Best Papers* **6**, No. 4, 511–536 (1997).
14. D. Wettschereck and T. G. Dietterich, "An Experimental Comparison of the Nearest Neighbor and Nearest Hyperrectangle Algorithms," *Machine Learning* **19**, No. 1, 5–27 (1995).
15. I. Jurisica, J. I. Glasgow, and J. Mylopoulos, "Incremental Iterative Retrieval and Browsing for Efficient Conversational CBR Systems," *International Journal of Applied Intelligence* **12**, No. 3, 251–268 (2000).
16. B. D'Ambrosio, "Process, Structure, and Modularity in Reasoning with Uncertainty," *Uncertainty in Artificial Intelligence 4*, R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, Editors, North-Holland Publishing Co., Amsterdam (1990), pp. 15–25.
17. E. J. Horvitz, "Reasoning under Varying and Uncertain Resource Constraints," *The Fifteenth National Conference on Artificial Intelligence AAAI'88*, Madison, WI, 1988, AAAI Press, Menlo Park, CA (1998), pp. 111–116.
18. I. H. Witten and E. Frank, *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Francisco, CA (2000).
19. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine* **17**, No. 3, 37–54 (1996).
20. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," *Journal of Molecular Biology* **247**, 536–540 (1995).
21. See http://bioinfo5.mbb.yale.edu/spine/sum.php3.
22. R. G. Farr, A. L. Perryman, and C. T. Samudzi, "Re-Clustering the Database for Crystallization of Macromolecules," *Journal of Crystal Growth* **183**, No. 4, 653–668 (1998).
23. C. L. Samudzi, M. J. Fivash, and J. M. Rosenberg, "Cluster Analysis of the Biological Macromolecule Crystallization Database," *Journal of Crystal Growth* **123**, 47–58 (1992).
24. D. Hennessy, V. Gopalakrishnan, and B. G. Buchanan, "Induction of Rules for Biological Macromolecule Crystallization," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* (*ISMB'94*), AAAI Press (1994), pp. 179–187.
25. D. Hennessy, B. Buchanan, D. Subramanian, P. A. Wilkosz, and J. M. Rosenberg, "Statistical Methods for the Objective Design of Screening Procedures for Macromolecular Crystallization," *Acta Crystallographica, D, Biological Crystallography* **56,** No. 7, 817–827 (2000).
26. B. Leng, B. G. Buchanan, and H. B. Nicholas, "Protein Secondary Structure Prediction Using Two-Level Case-Based Reasoning," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* (*ISMB'93*), AAAI Press, Menlo Park, CA (1993), pp. 251–259.
27. J. I. Glasgow, E. Steeg, and S. Fortier, "Motif Discovery in Protein Structure Databases," *Pattern Discovery in Molecular Biology: Tools, Techniques, and Applications*, Wang, Shapiro, and Shasha, Editors, Oxford University Press, New York (1999), pp. 77–95.
28. I. Rigoutsos, A. Floratos, C. Ouzounis, Y. Gao, and L. Parida,

"Dictionary Building via Unsupervised Hierarchical Motif Discovery in the Sequence Space of Natural Proteins," *Proteins* **37**, No. 2, 264–277 (1999).
29. P. Finn, S. Muggleton, D. Page, and A. Srinivasan, "Pharmacophore Discovery Using the Inductive Logic Programming System PROGOL," *Machine Learning* **30**, No. 2–3, 241–270 (1998).
30. J. I. Glasgow, "Array Theory and Knowledge Representation," *Arrays, Functional Languages and Parallel Systems*, L. Mullin, Editor, Kluwer Academic Publishers, New York (1992).
31. D. Conklin, S. Fortier, and J. I. Glasgow, "Representations for Discovery of Protein Motifs," *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, L. Hunter, D. Searls, and J. Shavlik, Editors, AAAI/MIT Press, Menlo Park, CA (1993), pp. 101–108.

Accepted for publication December 11, 2000.

Igor Jurisica *Ontario Cancer Institute, Princess Margaret Hospital, University Health Network, Division of Cancer Informatics, 610 University Avenue, Room 8-413, Toronto, Ontario M5G 2M9, Canada (electronic mail: ij@uhnres.utoronto.ca).* Dr. Jurisica is a scientist at the Ontario Cancer Institute, Princess Margaret Hospital, Division of Cancer Informatics. He holds adjunct professor positions at the Department of Computer Science, University of Toronto, and the Department of Computing and Information Science, Queen's University. He also holds a visiting scientist position at the Toronto Laboratory of the Centre for Advanced Studies, IBM Canada. Dr. Jurisica received a Ph.D. degree (1998) and an M.Sc. degree (1993) in computer science from the University of Toronto, and an M.Sc. degree (1991) in electrical engineering from the Slovak Technical University. His research interests include computational biology, case-based reasoning, machine learning, and knowledge discovery.

Patrick Rogers *Ontario Cancer Institute, Princess Margaret Hospital, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada (electronic mail: progers@alumni.uwaterloo.ca).* Mr. Rogers is a research assistant developing case-based reasoning tools for bioinformatics. He received his undergraduate degree in mathematics and computer science from the University of Waterloo in 1998. His interests include object-oriented operating system design, automated user interface design, and computer graphics.

Janice I. Glasgow *Department of Computing and Information Science, Queen's University, Kingston, Ontario K7L 3H2, Canada (electronic mail: janice@cs.queensu.ca).* Dr. Glasgow is Professor and Head of the Department of Computing and Information Science. She received her Ph.D. degree in 1983 at the University of Waterloo. Her research interests include computational imagery, molecular scene analysis, computational biology, and case-based reasoning.

Suzanne Fortier *Department of Chemistry, Queen's University, Kingston, Ontario K7L 3N6, Canada (electronic mail: fortiers@post.queensu.ca).* Dr. Fortier is a full professor in both the Department of Chemistry and the Department of Computing and Information Science at Queen's University. She is also vice principal (academic) at Queen's University and vice president of the Natural Sciences and Engineering Research Council (NSERC) of Canada. She received a Ph.D. degree in crystallography in 1976 from McGill University. Her research interests include crystallographic data mining and the development of methodologies for determining protein structure.

Joseph R. Luft *Hauptman-Woodward Medical Research Institute (HWI), 73 High Street, Buffalo, New York 14203-1196 (electronic mail: luft@hwi.buffalo.edu).* Mr. Luft is a research scientist at HWI and a member of the HWI Scientific Governance Council. He has been at the laboratory since receiving his B.A. degree in chemistry from D'Youville College in 1985. His research interests include the study of basic principles of macromolecular crystallization, development of new crystallization methods, and the application of high throughput technologies, imaging, and computational analysis to this work.

Jennifer R. Wolfley *Hauptman-Woodward Medical Research Institute (HWI), 73 High Street, Buffalo, New York 14203-1196 (electronic mail: wolfley@hwi.buffalo.edu).* Mrs. Wolfley is a research associate at HWI. She received a B.A. degree in chemistry and a B.S. degree in forensic chemistry from Buffalo State College in 1995. She worked on topical pharmaceuticals and drug delivery as an assistant research scientist at Bristol-Meyers Squibb from 1996 until 1999. She is currently working on high throughput crystallization method development at HWI. Her research interests include method development for macromolecular crystallization and the application of high throughput robotics and imaging to this work.

Melissa A. Bianca *Hauptman-Woodward Medical Research Institute (HWI), 73 High Street, Buffalo, New York 14203-1196.* Ms. Bianca will graduate from Geneseo College in May of 2001 with a B.S. degree in biology. She has been working at the laboratory as a research apprentice during semester breaks since 1996. She contributed toward the development of high throughput crystallization robotics at the laboratory and has worked on protocol development for the measurement of ancillary data.

Daniel R. Weeks *Hauptman-Woodward Medical Research Institute (HWI), 73 High Street, Buffalo, New York 14203-1196.* Mr. Weeks graduated from Geneseo College in 2000 with a B.S. degree in biochemistry and is currently earning a doctorate in pharmacology at the University of Buffalo. He worked at the laboratory as a research apprentice during semester breaks from 1998 until 2000. He contributed toward the development of high throughput crystallization robotics at the laboratory with his work on precipitation cocktail solutions.

George T. DeTitta *Hauptman-Woodward Medical Research Institute (HWI), 73 High Street, Buffalo, New York 14203-1196 (electronic mail: detitta@hwi.buffalo.edu).* Dr. DeTitta currently serves as Executive Director and Chief Executive Officer of the Hauptman-Woodward Institute. He joined the staff of HWI in 1970 after earning a B.S. degree in chemistry at Villanova and a joint Ph.D. degree in biochemistry and crystallography from the University of Pittsburgh. During his 30-year tenure at HWI, the topics of Dr. DeTitta's research have included direct methods of structure determination, biotin and prostaglandin structures, crystallographic programming, structural crystallography, and the macromolecular crystallization problem. His current research interests include the study of macromolecular crystallization and the phase problem in X-ray crystallography.