# Knowledge Management in Scientific Domains

**I. Jurisica**
University of Toronto
Toronto, ON M5S 3H5

**G. DeTitta** and **J. Luft**
Haumptman-Woodward Med. Res. Inst.
Buffalo, NY 14203

**J. Glasgow** and **S. Fortier**
Queen's University
Kingston, ON K7L 3N6

## Abstract

Scientific domains are characterized by substantial amounts of complex data, many unknowns, lack of complete theories, and rapid evolution. In many areas of decision making, much of the reasoning process is based on experience rather than on general knowledge. Experts remember positive cases for possible reuse of solutions, but negative cases are also useful for avoiding potentially unsuccessful results. Thus, storing and reasoning with experiences facilitates efficient and effective knowledge management.

In general, knowledge management systems support representation, organization, acquisition, creation, usage, and evolution of knowledge in its many forms. Complex scientific domains require: (1) multimodal representations that support application-domain richness, expressibility and domain-knowledge evolution, (2) effective organization of knowledge for efficient access to information, and (3) decision-support and analysis tools. We show how a case-based reasoning system can be used for knowledge management in structural biology. Namely, we describe a multimodal and multimedia system for managing crystallization experiences.

## Introduction

Biological research is generating data at an explosive rate. The Human Genome Project is expected to identify the codes for over 3 billion bases by the year 2005. This will provide code for about 100,000 proteins. About 350 folds and 1,200 super families have been studied experimentally. It is estimated that of the order of 1,000 folds and 3,000 – 5,000 super families still need to be studied (PSI 1998).

Analyzing this volume of data and using it intelligently is a challenge because of its complexity, its multiple interdependent factors, the uncertainty of these dependencies, and the continuous evolution of our understanding of the data. In general, reasoning with biomedical information requires flexible knowledge representation structures. Since knowledge in such domains is dynamic and evolutionary, its representation should support reasoning with a mix of relevant and irrelevant knowledge sources, allowing the scientist to deal with conflicting information. These domains often require the use of numeric, symbolic and multimedia information.

In this article we introduce a multimodal system, called MAX, for supporting the management of protein crystallization experience. We describe the design of knowledge repository as well as reasoning and analysis tools.

Problems arise in biomedical domains because information is not consistently described, quality control is not always in place across different laboratories, and often only positive results are reported. In order to support a systematic management of complex biomedical information and knowledge, we must resort to knowledge management techniques. Traditionally, database management systems, data warehouses and knowledge discovery in databases have been used to manage and use data. Data comprise values for observable, measurable or calculable attributes. Data in context is information. Knowledge is validated information (Firestone 1999).

A data management system is a computer program for managing a persistent and self-descriptive repository of data. Analogously, knowledge management systems support representation, organization, acquisition, creation, usage, and evolution of knowledge in its many forms.

A data warehouse is a subject-oriented, integrated, time variant and non-volatile set of data that supports decision making (Inmon 1996). Similarly, a knowledge warehouse is a storage vehicle for knowledge.

Knowledge discovery in databases is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Frawley & Piatetsky-Shapiro 1991). We may apply similar methods to knowledge bases (Jurisica *et al.* 1998) to discover underlying principles or meta-knowledge.

We incorporate a case-based reasoning (CBR) in the development of our intelligent management system for crystallization. Cases capture problem-solving processes by representing episodes of these processes. This makes CBR a suitable paradigm not only in domains that can be easily formalized but in hard-to-formalize areas as well. Because knowledge acquisition is supported by learning from experience, CBR systems can

supplement weak domain models. This requires representation formalisms which preserve relations among cases and also among their parts. CBR can represent existing experiences, deal with exceptions and contradictory information, evolve representation over time, change the use of information at a later time, and be effectively integrated with other systems.

## Knowledge Management

The application of knowledge-based technology to biomedical domains presents many challenges. These challenges arise from the complexity of biomedical knowledge characterized by a large number of interdependent factors, from the uncertainty of dependencies, and from its constant evolution. It is imperative that biomedical decision-support systems specifically address these challenges.

Decision-support systems have previously been applied to several problem in the domain of structural biology, for example: to help to identify protein secondary structures (Leng, Buchanan, & Nicholas 1993); to assist in locating motifs (Glasgow, Steeg, & Fortier 1999), to find similarity between protein tertiary structures (Grindley *et al.* 1994); and to help during the initial stage of drug discovery (Finn *et al.* 1998).

Some of the main problems to be solved in the development of intelligent support systems are how to represent information, how to access it flexibly and efficiently (Finn *et al.* 1998), how to analyze it (Conklin, Fortier, & Glasgow 1993), and how to reason with it during decision making. Given the uncertainties present, the diversity of the representation formalism used, the complexity and amount of information present, and the evolution of domain knowledge, it is necessary that the information system that will assist decision-making in structural biology be flexible and scalable. The retrieval of biomedical knowledge is difficult because of its various forms, diversity of its locations, and its potential to be contradictory.

Biological domains require multimodal representation that supports expressibility and domain-knowledge evolution. Although diverse tools need to be used, case-based reasoning can be the core technology due to its potential flexibility in managing experience. Next we discuss how a case-based reasoning system can be used for the management of crystallization experiments.

### $\mathcal{TA}3$ Case-Based Reasoning System

We propose to consider the representation of biomedical knowledge using a combination of case-based, image-based, and rule-based approaches. Selecting a particular formalism may require a tradeoff between information expressibility that the formalism supports, and scalability of the system that uses the formalism. In addition, an effective knowledge representation formalism supports knowledge evolution (Jurisica *et al.* 1998). Our approach takes CBR as the core system for knowledge management and extends its functionality through integration with other techniques. This is in fact a combination of two research schools – one that builds models of human performance in science and the other that builds effective programs, which may be implausible in human terms (Valdez-Perez 1995).

$\mathcal{TA}3$ is a CBR system which uses a variable-context similarity-based retrieval algorithm and a flexible representation language (Jurisica & Glasgow 1997; Jurisica *et al.* 1998; Jurisica, Glasgow, & Mylopoulos 1999). Cases are represented as a collection of attribute-value pairs. Individual attributes are grouped into one or more categories. Categories bring additional structure to a case representation. This reduces the impact of irrelevant attributes on system performance by selectively using individual categories during matching. As a result, we get a more flexible reasoning system, a more comprehensible presentation of complex information, improved solution quality, and increased scalability.

We extend basic CBR functionality in $\mathcal{TA}3$ by providing: (1) image-based processing to extend expressibility (Glasgow & Jurisica 1998), (2) database techniques for case retrieval to support scalability (Jurisica, Glasgow, & Mylopoulos 1999), (3) knowledge-discovery techniques to support domain-knowledge evolution and system optimization (Jurisica *et al.* 1998). The knowledge discovery component is used for three purposes:

- $\mathcal{TA}3$ optimization: locating descriptors relevant for a given context and task, and organizing a case base into context-based clusters.
- Case base and domain knowledge evolution: adding descriptors to assist case discrimination during prediction and classification; removing redundant cases and descriptors; creating hierarchies of descriptors and their values; finding associations.
- Evidence-based reasoning: analyzing created clusters, hierarchies and associations to identify underlying principles in the domain.

### Management of Crystallization Experiences

Structural biology is an important component of biological research. A standard method for protein structure determination is single crystal X-ray diffraction. This process is often limited by the difficulty of growing single crystals suitable for diffraction. Reasons for this include the large number of parameters affecting the crystallization outcome (e.g., purity of macromolecules, intrinsic physico-chemical parameters, biochemical, biophysical and biological parameters), the unknown correlations between the variation of a parameter and the propensity for a given macromolecule to crystallize.

Although some advances have been made, the crystallization of macromolecules is still primarily empirical. Because of its unpredictability and high irreproducibility, it has been considered by some to be an art rather than a science (Ducruix & Giege 1992). Practical experience produces theories that are effective in

many settings. For example, Jancarik and Kim proposed a set of 48 crystallizing agents that are often successfully used during crystallization (Jancarik & Kim 1991). These agents were proposed based on empirical results. Thus, *"further work in the systematic study of the chemical and physical properties associated with crystallization should provide additional improvements in the ability to grow protein crystals"* (PSI 1998).

An additional problem is an historically non-systematic approach to knowledge acquisition – *"the history of experiments is not well known, because crystal growers do not monitor parameters."* (Ducruix & Giege 1992). For example, the Biological Macromolecular Crystallization Database (BMCD) stores data from published crystallization papers, including information about the macromolecule itself, the crystallization methods used, and the crystal data (Gilliland 1988). Unfortunately, negative results are not reported, and many crystallization experiments are not reproducible because of incomplete method descriptions, missing details, or erroneous data. Our recent literature review indicates that the BMCD is not being used in a strongly predictive fashion.

Our main goal is to support the management of protein crystallization experiences by creating a useful crystallization knowledge warehouse. Conceptually, we can divide the crystal-growth process into a primary phase, during which one searches for favorable initial crystallization conditions, and a secondary phase, during which the initial conditions are optimized. Usually, the primary phase is the more time consuming of the two. Our research focuses on identifying initial conditions favorable for crystal growth as quickly as possible. The proposed approach has the potential to significantly reduce the time spent looking for initial conditions. The results of our research may thus eliminate the primary bottleneck in modern structural biology.

We postulate that past experience can lead us to the identification of initial conditions favorable to crystallization. Faced with the challenge of crystallizing a new protein we suggest that successful recipes developed for similar proteins may be optimal starting points for the lab work. However, we are faced with a problem of quantitatively measuring the similarity of two proteins.

We hypothesize that solubility experiments can provide a quantitative measure of similarity between two proteins. Assume that two proteins react similarly when tested against a reasonably large set of precipitating agents. We suggest that crystallization strategies successfully employed for the one may be profitably applied to the other. Thus, we need to 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 successfully crystallized before. New crystallization challenges are then approached by the execution and analysis of a small set of precipitation reactions, followed by the identification of similar proteins and further analysis of the recipes successfully used to crystallize them. The precipitation reactions are designed to consume less than a milligram of protein, to take a few hours to set up, and to take around 12 hours to analyze.

Our goals are to create a knowledge repository, develop software tools to analyze the outcomes and suggest starting conditions for crystal growth experiments, to develop a protocol for the execution of a set of precipitation reactions, and to design protocols for the visual interpretation of the reaction outcomes. To be specific, we have designed the decision-support system to identify the following pieces of information for the crystal grower: (1) the crystallization method of choice; (2) the crystallizing agent of choice; (3) the optimal temperature; (4) the optimal pH; and (5) the approximate concentrations of all solutes required in the crystal growth medium.

Crystallization experiments contain experiential information, such as initial input information about the protein at the beginning of the experiment, the process of carrying out the experiment and the outcome of the experiment. Knowledge evolves regardless of our wishes due to environment changes, the emergence of new problems or modification of our goals, user model changes, and the evolution of our understanding of the surrounding world resulting from the discovery of new relationships and principles. A case representation must be flexible to handle these issues. CBR suggests a model for computational reasoning that incorporates problem solving, understanding and learning. We propose to incorporate CBR to perform two functions: *(1)* to suggest almost-right solutions to problems, which can be modified automatically (or by the expert user) to suit the new protein situation, and *(2)* to warn of potential errors or failures in a proposed experimental plan.

Max, the proposed system for crystallization experiment design is multimodal and multimedia. It incorporates diverse tools (knowledge discovery, image feature extraction, similarity-based retrieval), reasoning algorithms (case-based, image-based and rule-based reasoning), and an information repository (multiple databases and a knowledge base). On the conceptual level, the information repository contains data and knowledge. Data comprises existing databases, such as PDB (Protein Data Bank), BMCD, and specialized information about proteins, chemicals, and agents. Knowledge in MAX's repository has two forms – experiential (cases in the case base) and general principles (e.g., adaptation rules). MAX's knowledge repository is created systematically with an emphasis on information quality (i.e., correctness, completeness, reproducibility). The case base stores cases, which are individual experiments with diverse crystallization outcomes (e.g., nothing happened, amorphous precipitate, crystalline precipitate, prisms). General principles can be useful rules acquired from crystallographers, or principles derived using knowledge-mining tools. These are needed during the case adaptation process.

The need for image-based reasoning comes from the

fact that there is no general solution to quantitatively evaluate reaction outcomes under the microscope. The major weakness of existing scoring methods is the tendency to confuse micro crystalline and amorphous precipitates (Ducruix & Giege 1992). For this reason we store crystallization outcomes as images. We use computer vision techniques (Glasgow & Jurisica 1998) to preprocess images (i.e., image alignment, recognizing precipitates in the micro pipette, etc.), to automatically recognize different crystallization outcomes, and to extract important image features for further analysis. It is important to note that this approach produces objective results, and thus may have the potential to be useful during knowledge discovery.

### Current Implementation Status of MAX

To support scalability, we have redesigned and reimplemented $\mathcal{TA}$**3**. The core of the library has been implemented in Java. We have extended the case management module to enable easier migration of cases, to support multimedia storage, and to improve scalability. Currently, $\mathcal{TA}$**3** works with both a memory-resident case repository and a JDBC-compliant database (e.g., IBM DB2 UDB).

We plan to evaluate individual components of MAX as we progress, but the final evaluation will be done after the information repository is populated and all reasoning and analysis tools implemented. The ultimate test is to take unknown proteins, use MAX to suggest crystallization strategies, and evaluate the results.

Using a limited data set, we have tested the suitability and accuracy of similarity-based retrieval using the solubility reaction index. A similarity function was used to determine which cases are most relevant to the given problem. Currently we incorporate precipitation indices as a quantitative measure of similarity. The index encodes precipitation reactions from a set of initial precipitation cocktails as a binary string. As the project develops we will also consider other attributes (such as protein sequence, molecular weight, etc.) and assess their usefulness in the retrieval stage of the system. The context-based retrieval method, implemented in the $\mathcal{TA}$**3** system, provides the user with a flexible interface for restricting or relaxing the context in order to retrieve fewer or more cases as necessary. The relaxation preferences for the reaction index category were initialized to favor reduction over generalization. Our preliminary results indicate that $\mathcal{TA}$**3**'s variable context, similarity-based retrieval module is suitable in protein crystallography domain. However, our evaluation is limited by the fact that reaction index was based on only 160 precipitation reactions performed. Currently, we are acquiring experimental data with a significantly larger number of precipitation reactions and for many more proteins.

The image processing system is divided into the retrieval and analysis modules. The retrieval module uses a simplified version of $\mathcal{TA}$**3**'s variable context retrieval. Each case is assumed to have only one attribute, which is a two-dimensional array of values. Contexts are represented by two arrays indicating a maximum and a minimum. Thus, the domain of a context is a continuous range of values. Contexts can be specialized and generalized, and sets of cases support regular and iterative retrieval, as well as an explain function. A set of cases is represented as a list of images resulting from a set of experiments.

The analysis module is used to extract features from the raw image to create a case. Image preprocessing isolates the region of interest in the image (i.e., the contents of the micro pipette) and attempts to standardize images with regards to lighting, size and orientation. Post-processing implements the classification strategy. While we investigate several possible strategies, currently only the analysis of the two-dimensional Fourier transform is implemented.

Although this is still only a prototypical version, our preliminary results are encouraging. We can now automatically clean up and align images, evaluate their similarity, and classify them into categories of "nothing happened" and "something happened" (e.g., amorphous precipitate, crystalline precipitate). We are working on extending the current implementation with additional algorithms for image feature extraction.

## Discussion

Successful knowledge management must be systematic. Without it, important data might be missing or incorrect, the representation formalism may be limited, or only a subset of available experience may be represented. Recently, other groups approached the problem of crystallography experiment design (Rosenberg *et al.* 1999). Similarly to our approach, they start by systematically archiving information about crystallization experiments, using machine-learning techniques to identify regularities in data, and suggest plausible crystallization experiments.

Traditionally, CBR focuses on technical domains or clinical medicine. Knowledge warehousing has been introduced in business-oriented domains (Firestone 1999). Our approach extends both CBR and knowledge warehousing techniques and applies them to scientific domains.

CBR is an important paradigm for knowledge management because: (a) it is similar to human problem solving and thus it can complement the user; (b) it supports evolving domain models and thus can help to increase domain understanding; and (c) it diminishes the problem of exceptions and over-generalizations.

Flexible similarity-based retrieval is required in complex application domains (Jurisica *et al.* 1998). It is also critical in biomedical domains, because similar structures are likely to have similar biological activities. It is, however, paramount that the case retrieval in such domains be scalable and user-guided, leading to conversational CBR (Munoz-Avila & Aha 1999). The prototype system $\mathcal{TA}$**3** can be used to satisfy these features (Jurisica & Glasgow 1997; Jurisica *et al.* 1998;

Jurisica, Glasgow, & Mylopoulos 1999). Further work is needed to extend the model to support reasoning about images, including automated extraction of image features and the assessment of geometric and spatial similarities (Glasgow & Jurisica 1998). Our preliminary results show that taking this multimodal approach to representation and reasoning has substantial benefits for knowledge management (Luft *et al.* 1999). Once the prototype becomes publicly accessible, additional problems with quality assurance will need to be addressed.

The unique aspect of this research is in its focus on domains where processing only symbolic information is not sufficient and where one representation formalism is not adequate to satisfy diverse users and support various tasks. The combination of a CBR paradigm with computer vision techniques may bring advances to decision support in this and similar domains. In addition, due to problem complexity, we focus on interactive rather than automatic tools. Results of this research are applicable beyond biomedical domains, provided experience-based reasoning is applicable. [1]

If biomedical decision-support systems (DSSs) can address these challenges, two important goals will simultaneously be achieved: (1) research in computing will be more focused and would progress faster since real world data is being used; and (2) research in biomedical domains will be enhanced by the application of new technology, which may help to establish underlying principles.

## References


Conklin, D.; Fortier, S.; and Glasgow, J. 1993. Knowledge discovery in molecular databases. *IEEE Transactions on Knowledge and Data Engineering* 5(6):985–987.

Ducruix, A., and Giege, R. 1992. *Crystallization of Nucleid Acids and Proteins. A Practical Approach.* New York: Oxford University Press.

Finn, P.; Muggleton, S.; Page, D.; and Srinivasan, A. 1998. Pharmacophore discovery using the inductive logic programming system progol. *Machine Learning* 30(2-3):241–270.

Firestone, J. M. 1999. Knowledge base management systems and the knowledge warehouse: A "strawman". http://km.org/AKMS/AKMSstrawbak.html.

Frawley, J., and Piatetsky-Shapiro, G. 1991. *Knowledge Discovery in Databases.* AAAI Press.

Gilliland, G. 1988. Biological macromolecule crystallization database. *Journal of Crystal Growth* 90(51).

Glasgow, J., and Jurisica, I. 1998. Integration of case-based and image-based reasoning. In Aha, D. W., ed., *AAAI'98 Workshop on Case-Based Reasoning*, 67–74. Madison, WI: AAAI Press.

Glasgow, J.; Steeg, E.; and Fortier, S. 1999. Motif discovery in protein structure database. In Wang; Shapiro; and Shasha., eds., *Pattern Discovery in molecular Biology: Tools, Techniques and Applications.* Oxford University Press. To appear.

Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; and Willett, P. 1994. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology* 229(3):707–721.

Inmon, W. H. 1996. *Building the data warehouse, 2nd Edition.* New York, NY: J. Wiley.

Jancarik, J., and Kim, S. H. 1991. Spare matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.* 24(409).

Jurisica, I., and Glasgow, J. 1997. Improving performance of case-based classification using context-based relevance. *International Journal of Artificial Intelligence Tools, Special Issue of ICTAI'96 Best Papers* 6(4):511–536.

Jurisica, I.; Mylopoulos, J.; Glasgow, J.; Shapiro, H.; and Casper, R. F. 1998. Case-based reasoning in IVF: Prediction and knowledge mining. *Artificial Intelligence in Medicine, Special Issue on CBR in Medicine* 12(1):1–24.

Jurisica, I.; Glasgow, J.; and Mylopoulos, J. 1999. Building efficient conversational CBR systems. Incremental iterative retrieval and browsing. Submitted.

Leng, B.; Buchanan, B. G.; and Nicholas, H. B. 1993. Protein secondary structure prediction using two-level case-based reasoning. In *ISMB'93.*

Luft, J. R.; Bianca, M.; Jurisica, I.; Rogers, P.; Glasgow, J.; Fortier, S.; and DeTitta, G. T. 1999. An opening strategy for macromolecular crystallization: Case-based reasoning and the exploitation of a precipitation reaction outcome database. In *Conference of the American Crystallography Association.*

Munoz-Avila, H., and Aha, D. W. 1999. Special issue on interactive case-based reasoning. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks and Complex Problem-Solving Technologies.*

PSI. 1998. National Institute of General Medical Sciences; Protein Structure Initiative meeting summary. http://www.nih.gov/nigms/news/reports/protein_structure.html.

Rosenberg, J. M.; Wilkosz, P. A.; Martowicz, M.; Chandeasekhar, K.; Subramanian, D.; Hennessy, D.; and Buchanan, B. 1999. Intelligent computation aids for crystal growth. In *Conference of the American Crystallography Association.*

Valdez-Perez, R. E. 1995. Computer science research on scientific discovery. Technical report, Carnegie Mellon University.



[1]This work is supported in part by grants from NASA and NSERC. The authors thank Melissa Bianca of the Hauptman-Woodward Medical Research Institute for invaluable help in the wet lab and Patrick Rogers of University of Toronto for invaluable contributions to the implementation of MAX.