AI Magazine Volume 25 Number 1 (2004) (© AAAI)

Applications of Case-Based Reasoning in Molecular Biology

Igor Jurisica and Janice Glasgow

■ Case-based reasoning (CBR) is a computational reasoning paradigm that involves the storage and retrieval of past experiences to solve novel problems. It is an approach that is particularly relevant in scientific domains, where there is a wealth of data but often a lack of theories or general principles. This article describes several CBR systems that have been developed to carry out planning, analysis, and prediction in the domain of molecular biology.

The domain of molecular biology can be characterized by substantial amounts of complex data, many unknowns, a lack of complete theories, and rapid evolution; reasoning is often based on experience rather than general knowledge. Experts remember positive experiences for possible reuse of solutions; negative experiences are used to avoid potentially unsuccessful outcomes. Similar to other scientific domains, problem solving in molecular biology can benefit from systematic knowledge management using techniques from AI. Case-based reasoning (CBR) is particularly applicable to this problem domain because it (1) supports rich and evolvable representation of experiences—problems, solutions, and feedback; (2) provides efficient and flexible ways to retrieve these experiences; and (3) applies analogical reasoning to solve novel problems.

CBR is a paradigm that involves solving new problems by recalling old problems and their solutions and adapting these previous experiences represented as cases. A case generally comprises an input problem, an output solution, and feedback in terms of an evaluation of the solution. CBR is founded on the premise that similar problems have similar solutions. Thus, one of the primary goals of a CBR system is to find the most similar, or most relevant, cases for new input problems. The effectiveness of CBR depends on the quality and quantity of cases in a case base. In some domains, even a small number of cases provide good solutions, but in other domains, an increased number of unique cases improves problem-solving capabilities of CBR systems because there are more experiences to draw on. However, larger case bases can also decrease the efficiency of a system. The reader can find detailed descriptions of the CBR process and systems in Kolodner (1993). More recent research directions are presented in Leake (1996), and practically oriented descriptions of CBR can be found in Bergman et al. (1999) and Watson (1997).

The remainder of this article describes several CBR systems that have been developed to address problems in molecular biology. We begin with a description of a recent CBR system for planning protein-crystallization experiments, followed by summaries of earlier CBR systems for gene finding, knowledge discovery in a sequence database, and protein- structure determination. We conclude with a discussion of issues related to the application of CBR in the domain.

CBR and Protein Crystallization

One of the fundamental challenges in modern molecular biology is the elucidation and understanding of the laws by which proteins adopt their three-dimensional structure. Proteins are involved in every biochemical process that maintains life in a living organism. Through an increased understanding of protein structure, we gain insight into the functions of these important molecules. Currently, the most powerful method for protein-structure determination is single-crystal X-ray diffraction, although new breakthroughs in nuclear magnetic resonance (NMR) (Kim and Szyperski 2003) and in silico (Bysrtoff and Shao 2002) approaches are growing in their importance. A crystallography experiment begins with a well-formed crystal that ideally diffracts X-rays to high resolution. For proteins, this process is often limited by the difficulty of growing crystals suitable for diffraction, which is partially the result of the large number of parameters affecting the crystallization outcome (such as purity of proteins, intrinsic physicochemical, biochemical, biophysical, and biological parameters) and the unknown correlations between the variation of a parameter and the propensity for a given macromolecule to crystallize.

The CBR system described in this section addresses the problem of planning for a novel protein-crystallization experiment. The potential impact of such a system is far reaching: Accelerating the process of crystallization implies an increased knowledge of protein structure, which is critical to medicine, drug design, and enzyme studies and to a more complete understanding of fundamental molecular biology.

Protein Crystallization

Conceptually, protein crystal growth can be divided into two phases: (1) search and (2) optimization. Approximate crystallization conditions are identified during the search phase, but the optimization phase varies these conditions to ultimately yield high-quality crystals. Improving these phases can lead to accelerated protein-structure determination and function resolution. Optimally, discovering the principles of crystal growth will eliminate protein crystallization as a bottleneck in modern structural biology.

The crystallization of macromolecules is currently primarily empirical. Because of its unpredictability and high irreproducibility, it has been considered by some to be an art rather than a science (Ducruix and Giege 1992) or an "exact art and subtle science." Experience alone has produced experimental protocols for crystal growth that are effective in many settings. For example, Jancarik and Kim (1991) proposed a set of 48 agents that are often used during crystallization. Although advances have been made through practical experience, a need remains for systematic and principled studies to improve our deep understanding of the crystallization process and provide a basis for the planning of successful new experiments. One of the difficulties in planning crystal growth experiments is that "the history of experiments is not well known, because crystal growers do not monitor parameters" (Ducruix and Giege 1992, p. 14). One recent approach attempted to optimize the 48-agent screen from crystallization data on 755 different proteins (Kimber et al. 2003). Not surprisingly, the study showed that one can eliminate certain conditions (in this case, 9) and still not lose any crystal or use only 6 conditions and still detect crystals for 338 proteins (60.6 percent). These results are encouraging. However, proteins have different properties, and thus, the selection of more or less useful crystallizing conditions might never be universal across all proteins. In addition, when a particular condition produces differential results across many proteins, it still might provide valuable information.

The BIOLOGICAL MACROMOLECULAR CRYSTALLIZATION DATABASE (BMCD) (Gilliland, Tung, and Ladner 2002) stores data from published crystallization papers (positive examples of successful plans for growing crystals), including information about the macromolecule itself, the crystallization methods used, and the crystal data. Attempts have been made to apply statistical and machine learning techniques to the BMCD. These efforts include approaches that use cluster analysis (Farr, Peryman, and Samuzdi 1998), inductive learning (Hennessy et al. 1994), and correlation analysis combined with a Bayesian technique (Hennessy et al. 2000) to extract knowledge from the database. Previous studies were limited because negative results are not reported in the database and because many crystallization experiments are not reproducible because of an incomplete method description, missing details, or erroneous data (which is the result of often skimpy and vague primary literature). Consequently, the BMCD is not currently being used in a strongly predictive fashion.

Significant advancement in this field has resulted from the use of high-throughput robotic setups for the search phase of crystal growth. This advancement vastly increases the number of conditions that can initially be tested and also improves reliability and systematicity for approximating conditions for crystallization in

the search phase. However, it also results in thousands of initial crystallization experiments carried out daily that require expert evaluation based on visual criteria. Thus, our CBR system includes an image analysis subsystem that is used to automatically classify the initial outcomes in the search phase. This classification forms the basis for the similarity measure for CBR. We incorporate knowledge discovery tools to assist in the optimization and the understanding of the protein-crystallization process.

Case-Based Reasoning System

One can view the process of crystal growth 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. Planning in AI generally involves developing techniques for forming a strategy of actions by choosing among alternative actions and reasoning about their effects. In CBR, actions are chosen based on the retrieval and adaptation of previously constructed plans.

Our approach builds on a previously developed computational framework for CBR called TA3. This system uses a variable context, a similarity-based retrieval algorithm, and a flexible representation language (Jurisica and Glasgow 2000; Jurisica, Glasgow, and Mylopoulos 2000). 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, reducing the impact of irrelevant attributes on system performance by selectively using individual categories during case retrieval.

Although we would hope that most pure proteins would crystallize readily under the appropriate conditions, finding a priori the optimal mix of solution conditions is challenging. One possible approach to make this search more systematic is the use of *factorial design*, which involves a set of experiments intended to identify important effects and interactions. In crystallography, it can be used to search for initial conditions and, once results are evaluated, to optimize the conditions. The process involves the simultaneous combination and modification of all conditions, which generally results in an intractable number of experiments.

We postulate that past experience can lead us to the identification of initial conditions favorable to crystallization. Moreover, it is believed that solubility experiments can provide a quantitative measure of similarity among proteins. If we consider only 15 possible conditions, each having 15 possible values, the result would be 4.3789e+017 possible experiments. Assume that 2 proteins react similarly when tested against a large set (over 1,500) of precipitating agents in the search phase of crystallization. Then, the planning strategies successfully used for the one can profitably be applied to the other. Thus, it is important to identify a suitable set of precipitating agents to sort the outcomes of reactions for a relatively large group of proteins. 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 an analysis of the recipes used to crystallize them (that is, crystal growth method, temperature and pH ranges, concentration of a protein, and crystallization agent). Figure 1 illustrates this process.

Our wet-lab collaborators, George DeTitta and others at the Hauptman-Woodward Medical Research Institute in Buffalo, New York, have the capacity to prepare and evaluate the results of more than 60,000 initial precipitation experiments, involving 40 proteins, during a single work week. These microbatch experiments are being conducted under paraffin oil using robots outfitted with syringes and cameras (figure 2). For each protein, crystal growth experiments are carried out using 1,536 different cocktails (solutions or reacting agents) (Luft et al. 2001). The result of a precipitation experiment is an image that is automatically generated and analyzed to determine the outcome of the crystallization process (Jurisica et al. 2001). Image processing involves four main steps: (1) drop recognition, (2) drop analysis, (3) image-feature extraction, and (4) image classification. A standard vocabulary of outcomes is utilized to describe the results of image classification: clear drop, amorphous precipitate, phase separation, microcrystal, crystal, and unknown. These outcomes, recorded as a function of time, are the cornerstone of the crystallization case base that also contains chemical and physical information about individual proteins.

A need for automated image analysis arises from the fact that there is no general approach to quantitatively evaluate crystallization reaction outcomes under the microscope. The major weakness of existing scoring methods is the tendency to confuse categories of precipitates (Ducruix and Giege 1992). In our approach, crystallization outcomes are stored as image representations that are analyzed using computer vision techniques (Jurisica et al. 2001) to objectively recognize different crystallization

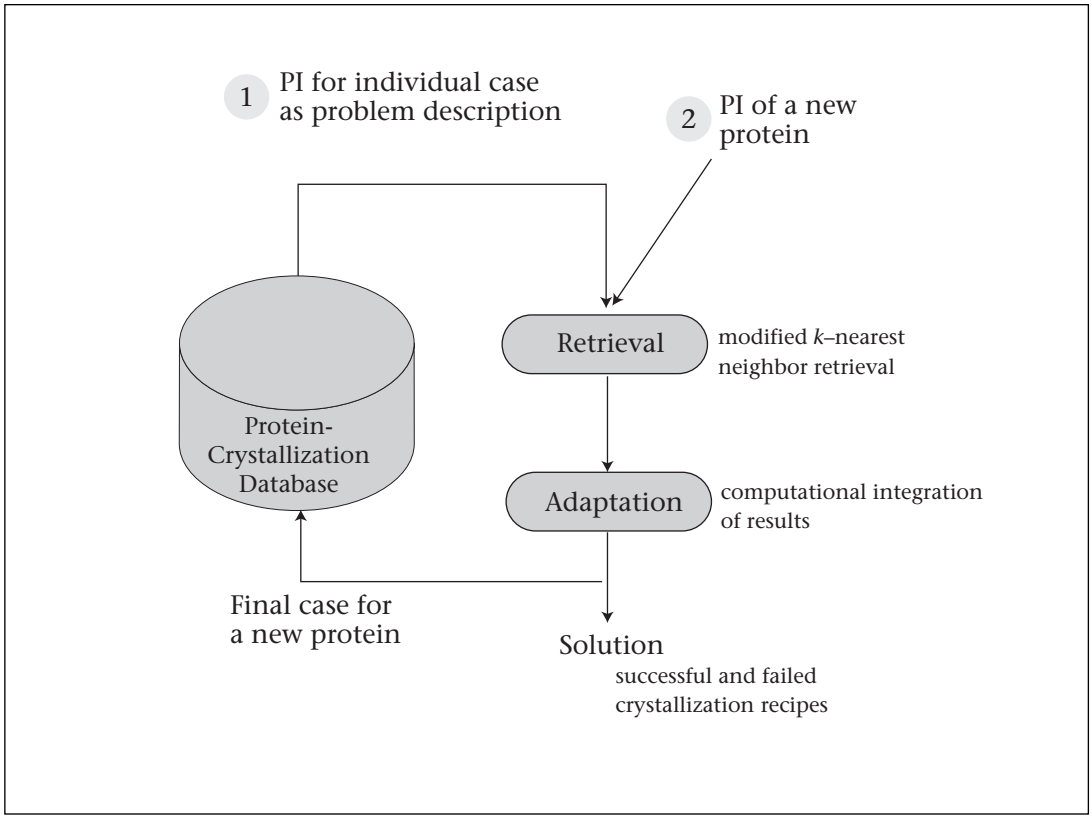

Figure 1. Case-Based Planning of Protein-Crystallization Experiments.

The precipitation index of a novel protein is compared to precipitation indexes of all proteins stored in the case base using a modified *k*–nearest-neighbor similarity matching. Crystallization plans of proteins with the most similar precipitation indexes are considered for planning a crystallization of a novel protein—successful crystallizations are used as positive planning experiences, but failed crystallization plans are utilized to potentially avoid negative results in the future.

outcomes and automatically extract important image features for further analysis. It is important to note that our approach produces objective results and is scalable to accommodate the vast number of images that require analysis. Our current database contains over 1,500 proteins, each being screened with 1,536 different conditions and photographed 6 times over a period of 2 weeks. Data storage thus requires approximately one DVD disk for each protein experiment.

Case retrieval involves partial pattern matching of an input case to cases in the case base. A similarity function is used to determine which cases are most relevant to the given problem. The precipitation index is utilized to define a quantitative similarity function for crystal growth experiments. This index encodes the reactions from the set of initial precipitation cocktails as a string containing the 1,536 initial crystallization outcomes (determined through image analysis). The distance is measured between the string for a novel set of reactions and those stored in the repository to extract those cases that are of minimum distance. The retrieval method implemented in the TA3 system provides the user with a flexible interface for restricting or relaxing the similarity function to retrieve fewer or more relevant cases as necessary (Jurisica, Glasgow, and Mylopoulos 2000).

Once similar cases have been retrieved, the next step in CBR is *adaptation,* which 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 appropriate parameters for a new set of experiments for protein crystallization. At this stage, the system acts, first, as an adviser to the crystallographer to suggest possible parameters for further experimentation and, second, as an evaluater of potential experiments that the user might propose. The system utilizes results from the precipitation index to suggest 80 percent of the solution, but 20 percent of the solution is determined from the knowledge obtained by data mining the repository for general trends. Although we still need to determine and vali-

date the most effective split, this combined approach to plan adaptation attempts to resolve the problem of local similarity versus global trends. The adaptation module is being developed over time as more general rules/principles are extracted from the growing case base using data-mining techniques. Once a plan (in the form of a set of experiments) has been derived and executed for a novel protein, the results are recorded as a new case that reflects this experience. Cases with both positive and negative outcomes are equally valuable for future decision-making processes and are also required for the application of data-mining techniques to the case base.

Currently, the system is being integrated, and preliminary validation is being conducted. The first step is image analysis, where we have processed almost 1,600 proteins to date (each screened on 1,536 different conditions). We have validated our automated image classification system, comparing it to human expert performance on 18 proteins, which resulted in an accuracy of 89 percent (receiver operating characteristic [ROC] score 0.875, precision 0.40, and recall 0.70) (Cumbaa et al. 2003). Some classification outcomes are illustrated in figure 3.

In automatic image classification, we use a boosting approach by combining results from multiple techniques. Our image analysis system comprises several stages: (1) well registration, (2) droplet segmentation, (3) feature extraction, and (4) image classification.

During well registration, we locate and eliminate the boundaries of the well. To determine the vertical-horizontal well boundaries, we find the pair of pixel columns/rows separated by the expected well width/height with the closest average pixel intensity. We incorporate a previously described registration method (Jurisica et al. 2001).

During droplet segmentation, we eliminate the edges of the drop. Currently, we use a probabilistic graphic model to segment the central region of the well for further analysis by distinguishing the empty well, the inside of a droplet, and the edge of a droplet. We first divide the well into a grid of 17×17 regions. Each region *(i,j)* in this grid is represented by a pair of variables in the Bayesian net segmentation model, which forms a 60-component mixture of multivariate Gaussian distributions. Each mixture models the local state of the well. Segmentation of a particular image is accomplished by inferring the most likely configuration from the probability distribution, using the local region of the image as well as the values of its neighbors. We trained our segmenta-

Figure 2. Crystallization Solution and Protein Pipetting Robots and XY Photography Stands.

The base set of robots in HWI laboratory comprises a liquid handling system pipette robot, which accommodates exchangeable banks of 1, 12, 96, or 384 syringes, that is utilized to deliver crystallization solution and protein into 1,536 wells (Robbins Scientific Tango, Sunnyvale, California) and a protein pipette robot with 96 syringes (Robbins Scientific Tango) and two XY stands for photographing experiment results, with the capacity to digitally photograph 86,016 crystallization experiments each.

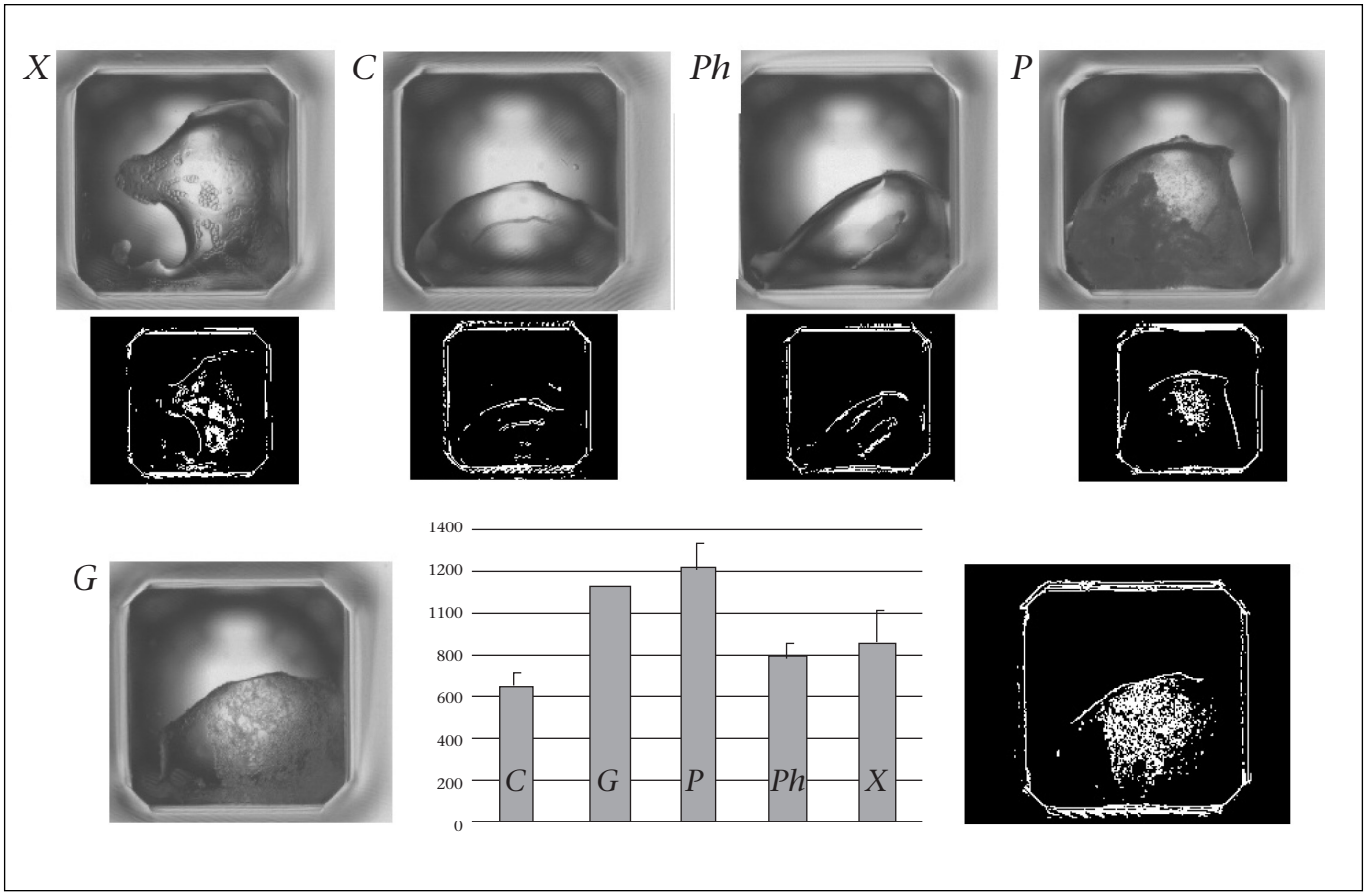

Figure 3. Automated Image Analysis and Classification.

After image segmentation, multiple classes of crystallization results can be detected. We use a boosting approach that combines multiple approaches. Here, we show images that have been classified as crystal (*X*), phase separation (*Ph*), precipitate (*P*), clear (*C*), and gel (*G*). The corresponding contour images have been utilized to compute the Euler number, which, in turn, has been utilized to cluster similar images. As shown in the bar graph, crystal and phase separation overlap but can be separated from clear drops and precipitates. Unfortunately, we did not have enough gel images to enable statistical analysis.

tion model on 45 hand-segmented images. Validation of the model was performed on a set of 50 hand-segmented images containing 4,319 empty-well regions, 1,348 droplet-border regions, and 8,783 intradroplet regions. The model correctly identified 96 percent of all well regions, 69 percent of all border regions, and 95 percent of all intradroplet regions, for a weighted mean of 93-percent overall accuracy (Cumbaa et al. 2003).

During feature extraction, we extract a minimal set of descriptive features that are essential in differentiating protein-crystallization outcomes. Because we approach the image-classification problem using an elimination strategy, we extract features that are indicative of specific subgroups of crystals (for example, microcrystals, microneedles, needle crystals, and larger faceted crystals), precipitates, and clear drops. In total, we reduce each image to a vector of 23 features: 20 measuring micro-crystal features, 2 measuring the presence of straight edges detected within a droplet, and 1 measuring the smoothness of the droplet content to detect precipitates and clear drops (Cumbaa et al. 2003).

During image classification, we classify each image, given its 23 numeric features, using *linear discriminant analysis,* into crystal, clear, and precipitate classes (Cumbaa et al. 2003).

Results of image classification have been utilized to effectively visualize time dependency of the crystallization process and begin the process of crystallization plan optimization. Considering that each protein crystallization results in 9,216 images, the main goal of the visualization system is to provide easier navigation through experimental results. Different size plates can easily be visualized by selecting a protein from a list and displaying a color-coded image classification across all time points (figure 4). Selecting a specific well displays cocktail information. Crystallization results are automatically (or interactively) selected for op-

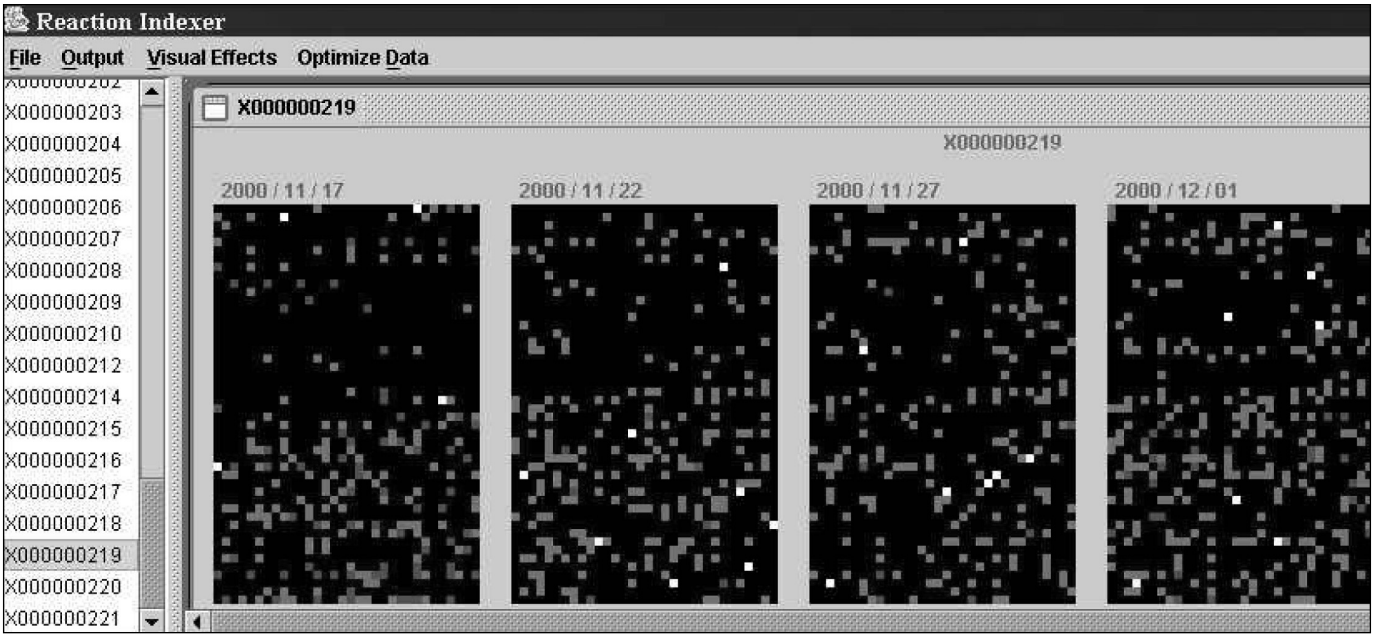

Figure 4. Visualization and Optimization of the Initial Crystallization Plan.

White squares denote crystals, light grey denotes precipitates, black indicates clear drops, and dark gray indicates unknown.

timization, using selected crystallizing conditions and optimization criteria (figure 5). At this stage, the cocktail setup is optimized into 24-well plate-optimization screen (figure 6). We combine information from the successful and failed crystallizations in a given screen with historical patterns obtained from data mining to plan novel experiments. This effort can be described as a combination of traditional crystallization optimization techniques with machine learning approaches.

Little research has been carried out in the area of AI and protein crystallization. As pointed out earlier, much of this research has focused on applying statistical and machine learning to the BMCD. Recent work on data mining applied to protein crystallization can be found in the article by Buchanan and Livingston contained in this issue of *AI Magazine*. Also worth noting is work by Hennessy et al. (2000), which describes an experiment planner for protein crystallization that uses a combination of CBR and Bayesian reasoning. In particular, they apply statistical methods to the BMCD to determine the probability of success of an experimental plan. Our approach differs from this work in several fundamental ways: (1) we are developing a case base that includes negative, as well as positive, experiments; (2) we are incorporating results from high-throughput robotic experiments to determine possible starting conditions for experimentation; (3) we are applying image-processing techniques to analyze experimental outcomes; and (4) we utilize the output of the image analysis to determine similarity between cases.

Other Applications of Case-Based Reasoning

Here, we summarize several other CBR systems that have been developed to address problems in molecular biology. In particular, we discuss approaches for analyzing genomic sequence data and predicting and determining the structure of proteins.

Sequence Analysis

The linear subsequences of deoxyribonucleic acid (DNA) that encode proteins are called *genes*. A three-letter string from the alphabet *A, G, T, C* (referred to as a *codon*) of bases from DNA encodes 1 of the 20 amino acids that are utilized to form a protein. One of the fundamental problems in the area of sequence analysis is *gene finding*, which involves locating the correct groups of nucleotide triples to translate into the protein sequence. This problem is complicated by noise in the data, which is caused by sequencing errors.

A CBR gene-finding algorithm was proposed by Jude Shavlik (1991). In this research, cases correspond to complete genes and complete proteins stored in existing biological sequence databases. *Similarity matching* is utilized to locate and identify regions in a DNA sequence that encode proteins, which is done in the

CocktailsX

Reactions	Well #	Cocktail #	Chemical Additive	Chemical Formula	Concentration	Buffer Type	Buffer Concentration	pH	MPD	MPD % (v/v)
	1225	2_C1225	Sodium nitrate	NaNO3	0.1 M	HEPES	0.1 M	7.5	MPD	40

Reactions	Well #	Cocktail #	Chemical Additive	Chemical Formula	Concentration	Buffer Type	Buffer Concentration	pH	PEG	PEG % (w/v)
	476	2_C0476	Calcium acetate	Ca(C2H3O2)2	0.1 M	MES	0.1 M	6	8000	20

Reactions	Well #	Cocktail #	Chemical Additive	Chemical Formula	Concentration	Buffer Type	Buffer Concentration	pH
	124	2_C0124	Potassium bromide	KBr	2.66M	MOPS	0.1 M	7

Figure 5. Selected Experiments for the Optimization.

One can identify both the time and the dependency of crystallization as well as view the details about crystallizing conditions.

Optimized pH Cocktails

File

Time (days) / Reacti...	Well #	Cocktail #	Chemical Additive	Chemical Formula	Concentration	Buffer Type	Buffer Concentration	pH
Fri Dec 01 20:00:0...	85	2_C0085	Magnesium acet...	Mg(C2H3O2)2*4...	1.79M	Tris	0.1 M	4
Fri Dec 01 20:00:0...	85	2_C0085	Magnesium acet...	Mg(C2H3O2)2*4...	1.79M	Tris	0.1 M	6
Fri Dec 01 20:00:0...	85	2_C0085	Magnesium acet...	Mg(C2H3O2)2*4...	1.79M	Tris	0.1 M	8
Fri Dec 01 20:00:0...	85	2_C0085	Magnesium acet...	Mg(C2H3O2)2*4...	1.79M	Tris	0.1 M	10
Fri Dec 01 20:00:0...	281	2_C0281	Magnesium nitrat...	Mg(NO3)2*6H2O	2.34M	Citrate	0.1 M	4
Fri Dec 01 20:00:0...	281	2_C0281	Magnesium nitrat...	Mg(NO3)2*6H2O	2.34M	Citrate	0.1 M	6
Fri Dec 01 20:00:0...	281	2_C0281	Magnesium nitrat...	Mg(NO3)2*6H2O	2.34M	Citrate	0.1 M	8
Fri Dec 01 20:00:0...	281	2_C0281	Magnesium nitrat...	Mg(NO3)2*6H2O	2.34M	Citrate	0.1 M	10

Time (days) / ...	Well #	Cocktail #	Chemical Addi...	Chemical For...	Concentration	Buffer Type	Buffer Concen...	pH	PEG	PEG % (w/v)
Fri Dec 01 20:...	431	2_C0431	Sodium molyb...	Na2MoO4*2H...	0.1 M	TAPS	0.1 M	4	20000	40
Fri Dec 01 20:...	431	2_C0431	Sodium molyb...	Na2MoO4*2H...	0.1 M	TAPS	0.1 M	6	20000	40
Fri Dec 01 20:...	431	2_C0431	Sodium molyb...	Na2MoO4*2H...	0.1 M	TAPS	0.1 M	8	20000	40
Fri Dec 01 20:...	431	2_C0431	Sodium molyb...	Na2MoO4*2H...	0.1 M	TAPS	0.1 M	10	20000	40
Fri Dec 01 20:...	474	2_C0474	Ammonium s...	(NH4)2SO4	0.1 M	HEPES	0.1 M	4	8000	20
Fri Dec 01 20:...	474	2_C0474	Ammonium s...	(NH4)2SO4	0.1 M	HEPES	0.1 M	7	8000	20
Fri Dec 01 20:...	474	2_C0474	Ammonium s...	(NH4)2SO4	0.1 M	HEPES	0.1 M	10	8000	20

Time (days) / ...	Well #	Cocktail #	Chemical Addi...	Chemical For...	Concentration	Buffer Type	Buffer Concen...	pH	PEG	PEG % (w/v)
Fri Dec 01 20:...	1136	2_C1136	Sodium molyb...	Na2MoO4*2H...	0.1 M	HEPES	0.1 M	4	400	80
Fri Dec 01 20:...	1136	2_C1136	Sodium molyb...	Na2MoO4*2H...	0.1 M	HEPES	0.1 M	7	400	80
Fri Dec 01 20:...	1136	2_C1136	Sodium molyb...	Na2MoO4*2H...	0.1 M	HEPES	0.1 M	10	400	80

Time (days) / ...	Well #	Cocktail #	Chemical Addi...	Chemical For...	Concentration	Buffer Type	Buffer Concen...	pH	MPD	MPD % (v/v)
Fri Dec 01 20:...	1301	2_C1301	Potassium ph...	KH2PO4	0.1 M	CAPS	0.1 M	4	MPD	20
Fri Dec 01 20:...	1301	2_C1301	Potassium ph...	KH2PO4	0.1 M	CAPS	0.1 M	7	MPD	20
Fri Dec 01 20:...	1301	2_C1301	Potassium ph...	KH2PO4	0.1 M	CAPS	0.1 M	10	MPD	20

Time (days) ...	Well #	Cocktail #	Commercial...	Chemical A...	Concentration	Chemical A...	Concentration	pH	Chemical A...	Concentration	Chemical A...	Concentrati
Fri Dec 01 ...	1444	2_C1444	HR Crystal ...	Ammoniu...	2.0 M	Tris Hydroc...	0.10 M	4	N/A	N/A	N/A	N/A
Fri Dec 01 ...	1444	2_C1444	HR Crystal ...	Ammoniu...	2.0 M	Tris Hydroc...	0.10 M	7	N/A	N/A	N/A	N/A
Fri Dec 01 ...	1444	2_C1444	HR Crystal ...	Ammoniu...	2.0 M	Tris Hydroc...	0.10 M	10	N/A	N/A	N/A	N/A

Figure 6. Twenty-Four Well Plate Optimization Screen for the Selected Conditions Using Specified Optimization Criteria.

presence of *frame shift errors*, that is, errors resulting from breaking the proteins into segments that result in different reading frames. The proposed approach involves parsing a noisy sequence into discrete cases that can be matched with entries in the case library. The gene-finding algorithm produces multiple, partial matches and then combines a subset of these into a consistent whole case. Aaronson et al. (1993) describe how CBR techniques can be applied to predict unknown regulatory regions in genes. The basis for cases in their system are instances of genes found in GENBANK (Benson et al. 2000), a primary repository for nucleic acid sequence data. An individual case corresponds to an abstraction over multiple gene instances. The prediction of regions is achieved using a grammatical model of gene structure. Indexing of cases is based on a hierarchy of attributes corresponding to protein and species similarity. Two approaches to prediction are proposed: (1) grammar-based CBR and (2) se-

quence-based CBR. In grammar-based CBR, instances of genes are represented in the form of *grammar rules,* which are applied to induce grammars of gene classes. These are then used to predict the presence and location of features for a novel sequence. Sequence-based CBR relies on the assumption that if there is similarity between two gene sequences, then these gene sequences might share similar features. Determining the location of a predicted feature in a novel sequence is determined by aligning the sequence with the similar sequence. The system was evaluated on how well it discovered features not in the feature table. It resulted in inferring 30 to 40 percent additional features that other existing extraction tools were not able to find.

Motivated by the observation of CBR in "think-aloud" protocols carried out by a human expert, Kettler and Darden (1993) utilized previous experimental plans, stored as cases, along with analogical reasoning, to plan for novel sequencing experiments. The planning component of their system created *plans*, which are sequences of laboratory and data analysis procedures. The primary objective of a plan is to find the amino acid sequence of a protein. A case in the knowledge base captures two types of experience: (1) *decision episodes,* which are the tasks that were chosen to be carried out, and (2) an *experiment episode,* which describes the experiment performed as a sequence of decision episodes. Case retrieval in the system consists of a simple probe and a similarity measure that counts the number of tasks that the current decision episode has in common with the stored cases. The approach is interactive, where the expert provides the system with potential hypotheses.

Protein-Structure Determination

The structure of proteins can be analyzed at varying levels of detail or complexity. The primary structure of a protein corresponds to an ordered chain of amino acid residues, or a *protein sequence*. Its secondary structure corresponds to the local conformation of the protein's backbone. The most common secondary structures are the α-helix and the β-pleated sheets. The *ternary structure* describes the unique three-dimensional arrangement of the atoms in the polypeptide chain of the protein. *Quaternary structure* describes the assembly of several individual polypeptide chains and, thus, is defined only for multipolypeptide proteins.

One way in which the tertiary structure of a protein can be represented is through the ϕ and ψ angles for each of the amino acids in the sequence. Zhang and Waltz (1989) describe a memory-based reasoning system to predict these angles for a protein based on known structures. Similar to CBR, memory-based reasoning makes use of specific past experiences for problem solving. Their work is based on the premise that if two amino acids have similar physical properties and occur in a similar environment, then they should have similar structure (in terms of their ϕ and ψ angles).

Leng, Buchanan, and Nicholas (1993) developed a CBR architecture for predicting the secondary structure of proteins. In this work, they considered several different measures from the biology literature for determining similar proteins. Once these proteins are found, their approach involves decomposing the novel sequence into smaller segments; retrieved cases are used to assign secondary structure to the corresponding piece of the unknown structure. Each amino acid in the sequence is assigned a class (α-helix, β-strand, or coil) by applying a weighted sum calculated from the evidence in the known structures.

CBR has also been applied to determine the three-dimensional structure of proteins from crystallographic data (Glasgow, Conklin, and Fortier 1993). This work in molecular scene analysis concerns the automated reconstruction and interpretation of protein image data (in the form of a three-dimensional electron density map). Cases correspond to previously determined structures. Discovered spatial and visual concepts of a structure are used to index cases. Cases are retrieved from the case base through a pattern-matching process that involves the comparison of unidentified features in a novel electron density map (derived from an image-segmentation process) with motifs from known structures. This approach combines a *bottom-up approach* to image analysis, where image-processing techniques are applied to extract features from the maps, with a *top-down approach,* where CBR is used to anticipate what motifs are likely to occur in the image.

Conclusions

CBR is based on the premise that problem solving involves recalling past experiences and utilizing these experiences (cases) to solve novel problems. It is a particularly useful paradigm in domains that are not well understood and where it is difficult to come up with generalizations that can be used to model the world.

The quantity and complexity of biological data being generated is increasing at a rate much faster than our ability to analyze or understand it, implying the need for advanced

computational tools for problem solving in the domain. This article has introduced several areas in which CBR has been applied to increase our understanding of biological data. However, a wealth of problems remains open that can benefit from AI in general, and CBR in particular. Issues that need to be explored to fully recognize the benefits of CBR for molecular biology problem solving include case representation and integrative CBR.

With *case representation*, most of the current biological databases have been designed for a specific single task and, thus, do not support the full potential of machine learning and CBR systems. For example, although the BMCD database can be considered as a case base of previous experiences in planning crystallization experiments, no negative experiences are stored, limiting its usefulness for CBR and machine learning.

With *integrative CBR,* applying CBR techniques to heterogeneous and distributed databases is increasingly required because relevant biological information can span multiple databases, where more than one source is necessary for problem solving. The vast amounts of genomic and proteomic data being generated require new analysis methods to combine, consolidate, and interpret the information. No single database or algorithm will be successful at solving the complex analytic problems. Thus, we need to integrate different tools and approaches, multiple sources of single types of data, and diverse data types. Similar trends have already been explored in the CBR literature, for example, McGinty and Smyth (2001) and Leake and Sooriamurthi (2002).

An advantage of CBR as a problem-solving paradigm is that it is applicable to a wide range of problems. It can be used to propose novel solutions or evaluate solutions to avoid potential problems. Aaronson, Juergen, and Overton (1993) suggest that analogical reasoning is particularly applicable to the biological domain, partly because biological systems are often *homologous* (rooted in evolution) and because biologists often use a form of reasoning similar to CBR, where experiments are designed and performed based on the similarity between features of a new system and those of known systems.

Acknowledgments

We would like to thank our collaborators George DeTitta and Joe Luft from the Hauptman-Woodward Medical Research Institute for stimulating long-term collaboration. We are also grateful to Nancy Fehrman and Angela Lauricella for the manual classification of a large number of training and testing images. Implementation of the presented software has been done by Christian Anders Cumbaa (image analysis), Patrick Rogers (CBR), and Xin Zhang (visualization and optimization). This research was supported in part by the Natural Science and Engineering Research Council of Canada, contracts 224114 and 203833; the National Institutes of Health, P50 GM62413; an IBM Shared University Research grant; and an IBM Faculty Partnership Award.

References

Aaronson, J. S.; Juergen, H.; and Overton G. C. 1993. Knowledge Discovery in GENBANK. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* eds. L. Hunter, D. Searls, and U. Shavlik, 3–11. Menlo Park, Calif.: AAAI Press.

Benson, D. A.; Karsch-Mizrachi, I.; Lipman, J.; Ostell, J.; Rapp, B. A.; and Wheeler, D. L. 2000. GENBANK. *Nucleic Acids Research* 28(1): 15–18.

Bergmann, R.; Breen, S.; Goker, M.; Manago, M.; and Wess, S. 1999. *Developing Industrial Case-Based Reasoning Applications: The INRECA Methodology.* Berlin: Springer-Verlag.

Bystroff, C., and Shao, Y. 2002. Fully Automated ab Initio Protein-Structure Prediction Using I-SITES, HMMSTR, and ROSETTA. *Bioinformatics* 18(Suppl 1): S54–S61.

Cumbaa, C.; Lauricella, A.; Fehrman, N.; Veatch, C.; Collins, R.; Luft, J.; DeTitta, G.; and Jurisica, I. 2003. Automatic Classification of Submicrolitre Protein-Crystallization Trials in 1536-Well Plates. *Acta Crystallographica Section D-Biological Crystallography* D59(9): 1619–1627.

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

Farr, R. G.; Perryman, A. L.; and Samuzdi, C. T. 1998. Reclustering the Database for Crystallization of Macromolecules. *Journal of Crystal Growth* 183(4): 653–668.

Gilliland, G. L.; Tung, M.; and Ladner, J. E. 2002. The Biological Macromolecule Crystallization Database: Crystallization Procedures and Strategies. *Acta Crystallographica Section D-Biological Crystallography* D58(1): 916–920.

Glasgow, J. I.; Conklin, D.; and Fortier, S. 1993. Case-Based Reasoning for Molecular Scene Analysis. In Case-Based Reasoning and Information Retrieval: Exploring Opportunities for Technology—Papers from the AAAI Spring Symposium, 53–62. Technical Report SS-93-07. Menlo Park, Calif.: AAAI Press.

Hennessy, D.; Buchanan, B.; Subramanian, D.; Wilkosz, P. A.; and Rosenberg, J. M. 2000. Statistical Methods for the Objective Design of Screening Procedures for Macromolecular Crystallization. *Acta Crystallographica Section D-Biological Crystallography* D56 (Pt 7): 817–827.

Hennessy, D.; Gopalakrishnan, V.; Buchanan, B. G.; Rosenberg, J. M.; and Subramanian, D. 1994. Induction of Rules for Biological Macromolecule Crystal-

lization. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology,* 179–187. Menlo Park, Calif.: AAAI Press.

Jancarik, J., and Kim, S. H. 1991. Spare Matrix Sampling: A Screening Method for Crystallization of Proteins. *Journal of Applied Crystallography* 24(409): 31–34.

Jurisica, I., and Glasgow, J. I. 2000. Improving Performance of Case-Based Classification Using Context-Based Relevance. *International Journal of Artificial Intelligence Tools (*Special Issue of IEEE ITCAI-96 Best Papers) 6(4): 511–536.

Jurisica, I.; Glasgow, J. I.; and Mylopoulos, J. 2000. Incremental Iterative Retrieval and Browsing for Efficient Conversational CBR Systems. *International Journal of Applied Intelligence* 12(3): 251–268.

Jurisica, I.; Rogers, P.; Glasgow, J. I.; Collins, R.; Wolfley, J.; Luft, J.; and DeTitta, G. 2001. Improving Objectivity and Scalability in Protein Crystallization: Integrating Image Analysis with Knowledge Discovery. *IEEE Intelligent Systems Journal* (Special Issue on Intelligent Systems in Biology) 16(6): 26–34.

Kettler, B., and Darden, L. 1993. Protein Sequencing Experiment Planning Using Analogy. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* 216–224. Menlo Park, Calif.: AAAI Press.

Kim, S., and Szyperski, T. 2003. GFT NMR, a New Approach to Rapidly Obtain Precise High-Dimensional NMR Spectral Information. *Journal of the American Chemical Society* 125(5): 1385–1393.

Kimber, M. S.; Vallee, F.; Houston, S.; Necakov, A.; Skarina, T.; Evdokimova, E.; Beasley, S.; Christendat, D.; Savchenko, A.; Arrowsmith, C. H.; Vedali, M.; Gerstein, M.; and Edwards, A. M. 2003. Data Mining Crystallization Databases: Knowledge-Based Approaches to Optimize Protein Crystal Screens. *Proteins: Structure, Function, and Genetics* 51(4): 562–568.

Kolodner, J. L. 1993. *Case-Based Reasoning*. San Francisco, Calif.: Morgan Kaufmann.

Leake, D. B., ed. 1996. *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. Menlo Park, Calif.: AAAI Press.

Leake, D. B., and Sooriamurthi, R. 2002. Automatically Selecting Strategies for Multi–Case-Based Reasoning. In *Advances in Case-Based Reasoning: Proceedings of the Fifth European Conference on Case-Based Reasoning,* 204–219. Berlin: Springer-Verlag.

Leng, B.; Buchanan, B. G.; and Nicholas, H. B. 1993. Protein Secondary Structure Prediction Using Two-Level Case-Based Reasoning. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* 251–259. Menlo Park, Calif.: AAAI Press.

Luft, J.; Wolfley, J.; Jurisica, I.; Glasgow, J. I.; Fortier, S.; and DeTitta, G. 2001. Macromolecular Crystallization in a High-Throughput Laboratory—The Search Phase. *Journal of Crystal Growth* 232(4): 591–595.

McGinty, L., and Smyth, B. 2001. Collaborative Case-Based Reasoning: Applications in Personalized Route Planning. In P*roceedings of the Fourth International Conference on Case-Based Reasoning,* 362–376. Berlin: Springer-Verlag.

Shavlik, J. 1991. Finding Genes by Case-Based Reasoning in the Presence of Noisy Case Boundaries. In *Proceedings of the 1991 DARPA Workshop on Case-Based Reasoning,* 327–338. San Francisco, Calif.: Morgan Kaufmann.

Watson, I. D. 1997. *Applying Case-Based Reasoning: Techniques for Enterprise Systems*. San Francisco, Calif.: Morgan Kaufmann.

Zhang, X., and Waltz, D. 1989. Protein-Structure Prediction Using Memory-Based Reasoning: A Case Study of Data Exploration. In *Proceedings of a Workshop on Case-Based Reasoning,* 351–355. San Francisco, Calif.: Morgan Kaufmann.

Igor Jurisica is an assistant professor in the departments of computer science and medical biophysics, University of Toronto, and the Department of the School of Computing, Queen's University. In addition to his position as a scientist at the Ontario Cancer Institute/Princess Margaret Hospital, Division of Cancer Informatics, Jirisica holds a visiting scientist position at the IBM Canada Toronto Laboratory. He is recognized for work in computational biology, including representation, analysis, and visualization of high-dimensional data generated by high-throughput biology experiments. His e-mail address is juris@cs.toronto.edu.

Janice Glasgow is a professor in the School of Computing at Queen's University, Canada, where she holds a research chair in biomedical computing. Currently, she is on sabbatical and is a senior visiting research fellow at the Institute of Advanced Studies, University of Bologna. She sits on the editorial board for several journals in AI, cognitive science, and bioinformatics; is a past-president of the Canadian Society for Computational Intelligence; and, until recently, was the vice-chair for the AI technical committee for the International Federation for Information Processing. Her e-mail address is janice@cs.queensu.ca.

ISMB Proceedings

ISMB-2000—San Diego, California

Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology

Edited by Russ Altman, Timothy Bailey, Philip Bourne, Michael Gribskov, Thomas Lengauer, Ilya Shindyalov, Lynn Ten Eyck, & Helge Weissig

ISBN 1-57735-115-0, 436 pp., index

ISMB-99—Heidelberg, Germany

Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology

Edited by Thomas Lengauer, Reinhard Schneider, Peer Bork, Douglas Brutlag, Janice Glasgow, Hans-Werner Mewes, & Ralf Zimmer

ISBN 1-57735-083-9, 324 pp., index

ISMB-98—Montréal, Quebec, Canada

Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology

Edited by Janice Glasgow, Tim Littlejohn, François Major, Richard Lathrop, David Sankoff, & Christoph Sensen

ISBN 1-57735-053-7, 234 pp., index

ISMB-97—Halkidiki, Greece

Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology

Edited by Terry Gaasterland, Peter Karp, Kevin Karplus, Christos Ouzounis, Chris Sander, & Alfonso Valencia

ISBN 1-57735-022-7, 382 pp., index

ISMB-96—St. Louis, Missouri

Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology

Edited by David J. States, Pankaj Agarwal, Terry Gaasterland, Lawrence Hunter, & Randall F. Smith

ISBN 1-57735-002-2, 274 pp., index

ISMB-95—Cambridge, England

Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology

Edited by Christopher Rawlings, Dominic Clark, Russ Altman, Lawrence Hunter, Thomas Lengauer, & Shoshana Wodak

ISBN 0-929280-83-0, 427 pp., index

ISMB-94—Stanford, California

Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology

Edited by Russ Altman, Douglas Brutlag, Peter Karp, Richard Lathrop, & David Searls

ISBN 0-929280-68-7, 401 pp., index

ISMB-93—Bethesda, Maryland

Proceedings of the First International Conference on Intelligent Systems for Molecular Biology

Edited by Lawrence Hunter, David Searls, & Jude Shavlik

ISBN 0-929280-47-4 468 pp., index

Published by The AAAI Press, 445 Burgess Drive, Menlo Park, California 94025 (www.aaai.org/Press)