The Crystallographer's Assistant

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

The only routinely used technique available today for obtaining the 3-D structure of a protein or DNA molecule is by X-ray diffracting a crystal of the macromolecule. The rate limiting step in structure determination is the process of growing a crystal of the macromolecule. This process is not very well understood, and can take a few weeks to several years. Crystallographers, therefore, are in great need of tools to aid them in the process of designing and performing experiments. There is a great deal of experiential data in this domain, in the form of scientific notebooks with graphical and textual representations of previous experiments.

We are in the process of collecting, analyzing and applying the knowledge available in this domain in order to design and develop the Crystallographer's Assistant (CA). The CA is an intelligent electronic assistant that will: help crystallographers record and maintain experimental context, offer suggestions as to experimental conditions that are likely to be successful for the current experiment (based on previously recorded successes and failures), and provide rationale for explaining failures (based upon theories that capture the significant relationships that exist in the data).

A set of about twenty-five parameters (e.g., pH, temperature) have been identified that affect the process of macromolecular crystallization[1]. Crystallographers systematically search this parameter space to find the optimal set of conditions under which a well diffracting crystal of the new macromolecule can be obtained. There exists only a preliminary understanding of the relationships that exist between two or more of these parameters.

In order to convince ourselves that it is indeed possible to find relationships among the various crystallization parameters from existing data, we have applied RL[2], an inductive learning program, to the data available in the Biological Macromolecular Crystallization Database (BMCD). The data in the BMCD is sparse, noisy, and represents only successful instances of crystal growth. In spite of the noisy nature of the data, RL has produced rules (and correlations) which have been considered significant by our domain experts. The limiting factor in the BMCD data is its lack of negative instances.

The Crystallographer's Assistant is based upon a case-based reasoning approach, and involves, as a first step, creating a database (from both existing experiment notebooks and on-going experiments), of about 1000 examples of crystallography experiments. These examples will provide us with both successful as well as failed experiments, and will be used both by RL as well as the case-based reasoner. Given the significant complexity and weak theory of the relationships between the features of the experiments, a case-based approach is being taken for similarity assessment. Experiential data concerning how the domain experts define pairs of cases to be similar and different will be used to guide the indexing and selection of cases. The result will be an experimenter's assistant which, given the results of the latest set of experiments, will remind the user of previous experiments with similar conditions and make suggestions based upon what was done in both cases of success and failure.

The results from applying RL to the BMCD data have yielded possibly significant new empirical relationships, as evaluated by our expert crystallographers. We are now in the process of applying RL to the newly created database of crystallography experiments. The next step will be to make the database available to researchers at other sites in order to expand the database to hold 100,000 or more cases. This will provide sufficient data to develop a more complete domain theory with the aid of modeling and machine learning techniques.

*Dr. Subramanian is affiliated with Department of Computer Science, Cornell University, Ithaca, NY 14853. This research is supported in part by funds from the W.M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University and the Pittsburgh Supercomputing Center.

References

[1] McPherson, A. Current approaches to macromolecular crystallization. *European Journal of Biochemistry*, 189 (1990), 1-23.

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