# Intelligent Aids for Parallel Experiment Planning and Macromolecular Crystallization

**Vanathi Gopalakrishnan** and **Bruce G. Buchanan** **John M. Rosenberg***
Intelligent Systems Laboratory
University of Pittsburgh
Pittsburgh, PA 15260
{vanathi, buchanan}@cs.pitt.edu
Fax:(412)624-9189 Phone:(412)624-9181

Department of Biological Sciences
University of Pittsburgh
Pittsburgh, PA 15260
jmr@jmr3.xtal.pitt.edu
Fax:(412)624-8109 Phone:(412)624-4636

## Abstract

This paper presents a framework called Parallel Experiment Planning (PEP) that is based on an abstraction of how experiments are performed in the domain of macromolecular crystallization. The goal in this domain is to obtain a good quality crystal of a protein or other macromolecule that can be X-ray diffracted to determine three-dimensional structure. This domain presents problems encountered in real-world situations, such as a parallel and dynamic environment, insufficient resources and expensive tasks. The PEP framework comprises of two types of components: (1) an information management system for keeping track of sets of experiments, resources and costs; and (2) knowledge-based methods for providing intelligent assistance to decision-making. The significance of the developed PEP framework is three-fold – (a) the framework can be used for PEP even without one of its major intelligent aids that simulates experiments, simply by collecting real experimental data; (b) the framework with a simulator can provide intelligent assistance for experiment design by utilizing existing domain theories; and (c) the framework can help provide strategic assessment of different types of parallel experimentation plans that involve different tradeoffs.
**Keywords:** Trial-and-error learning, simulation.

## Introduction

In some experimental science, a very large parameter space needs to be searched to find one or more sets of conditions that define a satisfactory solution. Such searches have to be performed taking into account resource limitations, which exist in almost all real-world problems. We have focussed our research on one such problem, namely, that of experiment design in macromolecular crystallization. In this domain, the goal of experimentation is to obtain a good quality (X-ray diffractible) crystal of a macromolecule (protein, DNA, or protein-DNA complexes). Such a crystal can be obtained only under certain specific conditions, which vary from macromolecule to macromolecule, that arise due to the mixing of different chemical compounds under varying conditions of physical factors such as temperature, pressure and gravity. There are some specific features related to growing crystals of macromolecules that make this problem particularly interesting. These include:

1. effects of actions changing over time, that is, partial results of experiments vary over time. For example, we could observe a crystal in an experimental apparatus on one day, but a week later it could have dissolved completely.

2. imprecise evaluation of partial results. There exists only a crude local evaluation function (based on visual inspection) for partial result determination of each experiment.

3. a large degree of interdependence among the variables, with the relationships of their interaction largely unknown.

4. the tedious nature of experimentation. Long hours are spent in repeatedly pipetting solutions into experimental apparatus since typically several (200 to 300) experiments are performed in parallel.

Tradeoffs are necessary in order to successfully search a large multi-dimensional parameter space to find a small region that yields a good quality crystal of a macromolecule.

Given these problem characteristics, one central question is the following: Can we infer global strategies for designing several experiments in parallel that can lead to a satisficing solution, given (a) limited resources with costs associated with use of each resource, (b) effort involved in performing each experiment,(c)

* Most often in AI, the term domain expert is used to describe a person with significant experience in the domain. We see the necessity to integrate rather than separate the varying domains, mainly due to the constant dialogue and transfer of knowledge between the parties involved. Hence, the use of *we* in this paper includes the people with significant experience in the domain.

partial results that vary over time and (d) only a crude local evaluation function (that provides partial result for an experiment at any given time of observation)? To rephrase the question from a crystallization viewpoint – what is a good strategy for crystallizing an unknown protein? Do we spend almost all of the protein we have on initial screening experiments, or use a small sample initially? The only "real" way to find out is to repeat the entire process using different approaches. That is not possible with real proteins due to unavailability of large amounts of protein and/or the costs involved with protein purification. Hence, there arises a need for the PEP system and simulator of hypothetical protein crystallization behavior, that are described in this paper.

This research describes a framework, called the PEP (Parallel Experiment Planning) framework, within which global strategies can be tried by a human designer. The framework includes a predictive model that can be used for simulating an experimental outcome at a particular time, and provides reflective statistical summaries of the choices made for experimentation. Different scenarios that are representative of the difference in response behavior of hypothetical proteins in various physico-chemical environments over time can be produced using simulation. The framework is general enough for use without a simulator - actual observations can be noted instead. The usefulness of the framework is greatly increased, however, with the use of simulation, since the implementation can be used as an intelligent decision-making aid and as a tool for training novice crystallographers to devise strategies for parallel experimentation for different types of macromolecules.

## Background and Motivation

Crystallization is an essential first step in macromolecular 3-D structure determination by X-ray crystallography. This is the only method capable of revealing high resolution structures for most proteins, protein-DNA complexes, viruses etc. The high resolution structural information is critical for modern molecular biological methods which rely on knowledge of the geometrical interrelationships of the various components that comprise the overall structure. (Multidimensional NMR can also determine macromolecular structures, but it is limited to molecules whose molecular weight is under 20,000; most proteins are larger than that).

The rate limiting step in X-ray structure determination is the crystallization itself. It takes anywhere between a few weeks to several years[1] to obtain macromolecular crystals that yield good diffraction patterns. The theory of forces that promote and maintain crystal growth is preliminary, and crystallographers systematically search a large parameter space of experimental settings to grow good crystals.

A set of about twenty-two parameters (such as temperature, pH, pressure, etc) that determine success in crystal formation have been empirically identified (Bergfors 1990). Crystallization attempts begin with the design of an initial screening experiment that coarsely samples the parameter space with a small number of parallel probes (typically between 200 and 300). The initial experiments usually incorporate only general information about crystallization combined with some specific details regarding the individual molecule under consideration (stability data, solubility data, isoelectric point, etc). One approach begins with a very coarse, uniform grid that spans the variables of interest (Weber 1991). Another begins with an incomplete factorial design that randomly samples the variables such that all pairwise combinations are tested (Carter & Carter 1979). The results from the initial screens are used to design finer sampling strategies for subsequent rounds of crystallization trials.

Rational parameter sampling in the iterated experimental protocol is made difficult by the need to run experiments over several months as well as by the nonlinear behavior of macromolecules with respect to the different parameters. The length of time between iterations makes it a challenge for humans to remember contextual information. Crystallographers presently rely on paper logs of experiments for this purpose; however the types of access they support are very limited. In order to facilitate the capture of experimental information electronically, an effort has been made to build an electronic laboratory notebook (Gopalakrishnan *et al.* 1994b; Hennessy *et al.* 1994). Even though the software XtalGrow is currently being used by about 10 laboratories in the United States, we are still faced with the problem of insufficient data for purposes of analysis. Also, the software is still evolving.

The main source for data about successful crystallizations is the Biological Macromolecule Crystallization Database (BMCD) (Gilliland 1987). This database captures information about successful experiments only. Several attempts have been made to exploit the available data in order to design initial crystallization trials for unknown macromolecules. Samudzi et al. (Samudzi, Fivash, & Rosenberg 1992) performed a cluster analysis on version 1.0 of the

[1] In fact , the question of how long to wait for crystals or precipitate to appear in a cell is not easy to answer. Anecdotes abound on experimenters finding crystals in plates abandoned for a year or more in the laboratory.

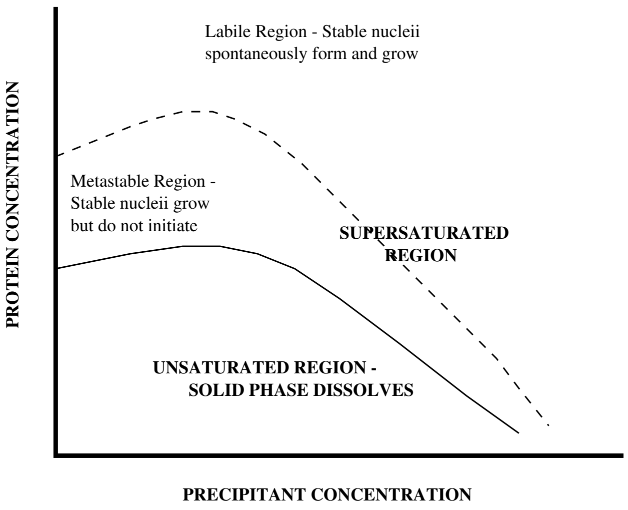


Figure 1: Two important dimensions of phase diagram for protein crystallization (from (Feigelson 1988))

BMCD and suggested a set of screening conditions specific to a major class of macromolecules. In Gopalakrishnan et al. (Gopalakrishnan *et al.* 1994a), preliminary attempts were made to recreate the clusters obtained in Samudzi et al. using two kinds of methods – statistical analysis (same as Samudzi's) and COBWEB (Michalski & Stepp 1983) (a machine learning and discovery program). The results from the clustering analysis were then used as input to the RL (Provost & Buchanan 1992) inductive rule-learning program, resulting in verification and expansion of Samudzi's results. Hennessy et al. (Hennessy *et al.* 1999) augmented the BMCD with a hierarchical classification of the macromolecules contained therein, as well as data on the additives used with them, and performed a statistical analysis that has led to a Bayesian technique for postulating the degree of success of a set of experimental conditions for a new macromolecule belonging to some known class.

## An Experiment in Macromolecular Crystallization

The basic crystallization experiment is to slowly reduce the solubility of a sample solution of the macromolecule by one of several established methods(Ducruix & Geige 1992). The solubility is determined by all the environmental parameters, one of which is usually the concentration of a "precipitating agent", such as polyethylene glycol (PEG), a commonly used precipitant. The "crystallization method", such as *vapor diffusion*, slowly raises the concentration of the precipitating agent (and almost everything else). If all the conditions are favorable, a point is reached where a crystal nucleates and grows. Figure 1 shows the interaction between two influential parameters for macromolecular crystal nucleation and growth, namely the concentrations of protein and precipitant. The supersaturation required for nucleii to form is much higher than that needed for growth. Thus, the ideal conditions for growing good quality crystals are generally considered to be along the boundary between the metastable and labile regions. Nucleii that form far into the supersaturated region are most likely to precipitate due to the fast rate of growth. The unsaturated regions indicate the unlikelihood of crystals forming and correspond to clear experimental results.

The basic experiment is repeated with different parameters until the experimenter succeeds, abandons the effort entirely, or decides to work on crystallizing a mutant or variation of the original macromolecule. Typically, many experiments (between 200 and 300) are started simultaneously and allowed to run for several weeks to several months. During this time, the experimenter attends to other projects. Then, the results are evaluated and a new series begun (to run concurrently with the older ones). Thus, large volumes of data accumulate over long periods of time. A slow rate of change is very often essential for crystal growth, thus there may be no other way to speed the process than to find good environmental parameters quickly.

An experiment can be described as containing values for three sets of variables, *givens*, *controllables* and *observables*. *Givens* represent the known information about a protein such as its identity, molecular weight, and isoelectric point. *Controllables* represent the values for the control parameters such as the concentrations of protein and precipitant, pH, and temperature. The term *observables* is used to denote the vector of partial results that are observed at different times. The givens determine the observable effects of the choice of controllables. As the observables change over time, they are referred to as partial results at any particular time of observation. This classification of the variables into givens, controllables and observables helps us understand the manner in which the different types of variables influence one another.

Figure 2 gives a pictorial description of how experiments and trials are done by hand. Each experiment is shown as a circle within a tray that can be used to set up a maximum of 24 experiments. Several trays are set up as the initial trial or set of parallel experiments. Each experiment includes some measures for control variables such as protein, precipitant and salt concentrations. Each well or experiment is examined under a microscope to yield observables over time. Depending on partial results observed, new trials are started and

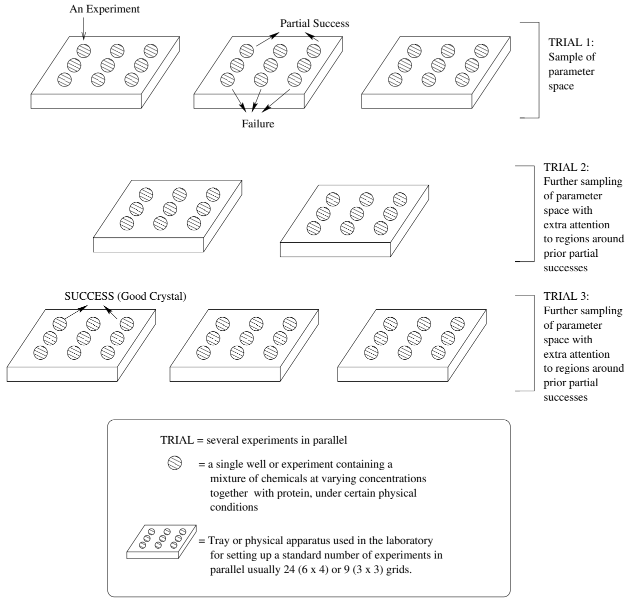


Figure 2: Parallel experimentation in macromolecular crystallization. After each trial is set up: success = good crystal, partial success = precipitate or unusable crystal, and failure = clear or no change in solution

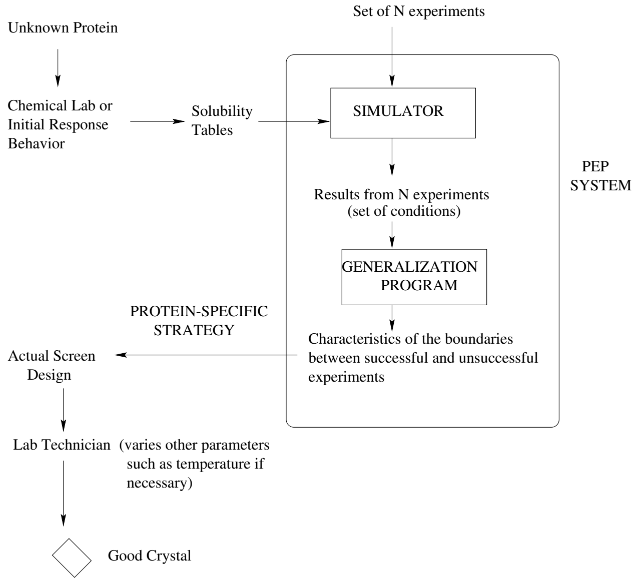


Figure 3: The general idea for providing intelligent assistance

examined, until a good quality crystal is obtained.

## Motivation for the Framework

The above description of how experiments are actually designed and performed in parallel forms the basis for the abstracted PEP framework. Because there are insufficient data regarding both successful and unsuccessful experiments (and the degree to which they are successful/unsuccessful), we needed a framework within which we could study the feasibility and requirements for capturing and analyzing experimental data. One commonly adopted method for such a study is modeling and simulation. This research includes the construction of an approximate physico-chemical model of crystal nucleation and growth. Most of the validation for the performance of the model was subjective, with the satisfaction of the domain experts as the major goal.

Figure 3 depicts the ideas that form the basic motivation for this research. When a crystallographer is faced with an unknown protein with only a few given information, such as its molecular weight and its isoelectric point or pI[2], the goal is to obtain at least one crystal that is good for X-ray diffraction as soon as possible and without running out of protein material.

A good quality crystal is usually one that has a resolution limit of diffraction (difflim) less than 3 Å. In special cases, there exists only a limited amount of protein (such as when a patient is dead, and is the only source of a particular protein). Also, protein isolation and purification is an expensive process. The major factors that limit the number of experiments that could be performed in the laboratory therefore are the amount of protein and the tedious nature of experimentation. Thus, any mechanism or method that would be able to help in cutting down the the number of experiments that we would try before getting at least one hit (good crystal), would be most helpful to crystallographers. The research undertaken herein attempts to lead us toward such methodology.

The method shown in Figure 3 attempts to use preliminary solubility data available for an unknown protein that needs to be crystallized, as input for a model that could be used to simulate experimental outcomes over time. The space of experiments defined within the simulator model consists of a reduced set of factors that could possibly influence the outcomes. If the assumptions that are made in the model hold in the real-world, we would then be able to predict which (sets of) experiments within this reduced space are more likely to yield a good quality crystal. These identified experimental conditions could then be further manipulated in the laboratory, if necessary, by varying parameters that are outside the scope of the simulator model. Thereby, using existing theory and a model based on such theory to predict the likelihoods of success for experiments

[2] At its isoelectric point, a macromolecule carries an equal number of positive and negative charges, and is therefore electrostatically neutral or balanced.

along some dimensions of the search space, we could reduce the total number of experiments that need to be performed in the laboratory before a single good crystal of any given protein is obtained.

In this research, we have constructed a fairly sophisticated simulator based on a predictive model of macromolecular crystallization. Given such a predictive model, we can try out a large number of experiments in virtual space to understand the characteristics of the boundary that separate the successful experimental conditions from the unsuccessful ones. Identification of the boundary between the classes of observable results is an important aspect of being able to come up with some internal model of how the given protein seems to be responding under different experimental controls. This gives some insight into possible rates of change of various hidden variables that influence crystal nucleation and growth such as the saturation level of protein in solution. A generalization program (called C4.5 (Quinlan 1993)) is used within the framework to provide a human-understandable description of the boundary based on all the experiments and partial results observed so far. By using the framework with the simulator and generalization program to learn such protein-specific strategies to try out in the laboratory, we hope that we will be able to reduce the number of actual experiments tried before obtaining a good quality crystal. Using this method we would be able to save time, effort and material involved in the crystallization of an unknown protein.

## The PEP Framework

The framework for designing parallel experiments as concurrent trials is shown in Figures 4 and 5. The framework has been developed to facilitate the setting up of parallel experiments as a defined region of the vast search space that is being searched at some level of granularity. Parallel experiment planning is therefore viewed as heuristic search with the human experiment designer as the heuristic control element. The control exercised by the human user is both appropriate and necessary in this domain.

The framework uses terms that have the following definitions:

1. *Experiment*: A combination of conditions (values for control parameters) that result in a series of outcomes at different times and consume some set of resources.

2. *Trial*: A group or population of experiments that are performed in parallel.

3. *Project*: the problem (e.g. the protein to be crystallized).

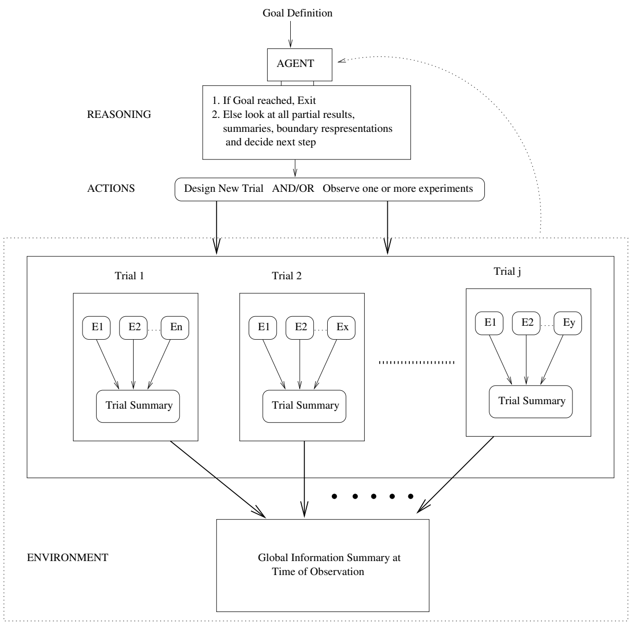


Figure 4: Framework for designing parallel experiments as concurrent trials

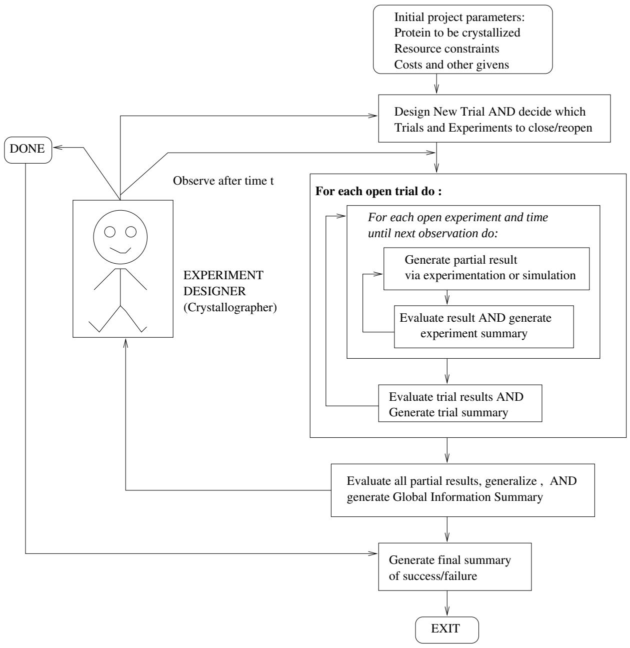


Figure 5: Details of one trial shown within the PEP system implementation

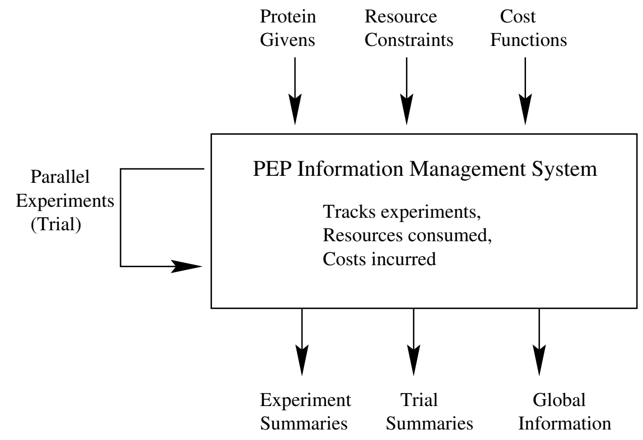


Figure 6: PEP information management system overview

4. *Global Information*: Information about the constants and variables for a project, as well as experimental record. This record consists of statistical summaries, and boundary information that represents the evolving boundary between classes of observable partial results.

5. *Local policy*[3] *(or search heuristic)*: the basis for defining a new trial after observing partial results.

6. *Global strategy (or search protocol)*: the abstract description of a series of trials leading to the goal (such as a good quality crystal).

If we were to try to represent the crystallization problem as an AI planning problem, it would have to be cast as a reactive planning problem, as the environment is dynamically changing. The biggest challenge would arise in how to (re)plan based on evaluation of partial results. Thus, we would need an evaluation function that can assign probabilities to different regions of the search space of experiments based on observation and evaluation of partial results over a period of time. These factors make this problem very different from traditional AI planning problems(Russell & Norvig 1995).

## Key Ideas

The main ideas that are represented in the framework in Figure 4 can be stated as follows:

1. Economic variables are important in strategic decision making, and hence must be included as part of every data gathering and data analysis project.

[3] A complete mapping from states to actions describes a policy. A policy therefore represents a simple reflex agent that knows what action to perform based on what state it is in.

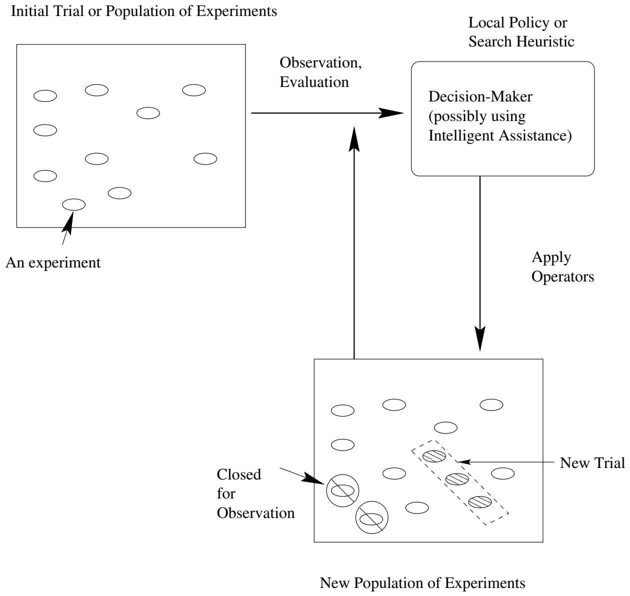


Figure 7: The PEP search

2. The nature of the real-world is such that multiple actions are taken in parallel, and the effects of some critical actions/decisions are visible only in the long-term, the evaluation is not precise as it is often based on visual information, and resources such as time and money are constrained.

3. In order for parallel experiments to be designed, it is necessary to design variables/data structures that can manage the information in an aggregate manner. Figure 6 depicts an overview of the PEP information management system that keeps track of experiments performed, resources consumed and costs incurred. The system maintains information that pertain to experiment, trial and global information summaries. Experiment summaries include information about the best partial result observed so far for each experiment, and the time of observation of the best result.

4. In order to provide intelligent assistance for decision-making, it is necessary to implement first the PEP information management system that can maintain the information about experiments in electronic form. Intelligent assistance can then be provided by building partial or approximate models, and using them to simulate experimental data for input to a generalization program. The generalization program can help identify the boundary characteristics that separate the successful and unsuccessful experiments given the model.

Figure 7 depicts an overview of the search involved in the PEP framework. An initial population of experiments represents the first trial. These experiments are observed and evaluated, and based on evaluation of these partial results, a decision-making agent applies operators to generate a new trial and close (or reopen) one or more experiments for observation. The initial set of experiments are still running concurrently. The simulator returns an observation for any experiment after a specified period of time (a real number representing number of weeks). The generalization program represents intelligent evaluation of experiments and their results over time, and provides a representation for boundary characteristics between observed classes of results in terms of rules containing a conjunction of control values on the left-hand side and the observable class (such as crystal) on the right-hand side. An example of a set of simple rules is shown below:

```
Rule 2:
        protconc > 1 ->  class 1 (CRYSTAL)

Rule 1:
        protconc <= 1 ->  class 3 (CLEAR)
```

The rules indicate that protein concentration was the most influential variable in providing the distinction between classes 3 (clear) and 1 (good crystal). Thus, using these rules, the experiment designer can carry out future trials, wherein protein concentrations less than or equal to 1 mg/ml are not used. For details regarding the representation and data structures used by the information management system of the PEP framework, see (Gopalakrishnan 1999).

## Descriptions of Main Parts

The main components of the framework that supports parallel experiment planning in this domain are discussed below.

**Design Trial** This is the first step or task that needs to be undertaken by an agent. Initially, only resource constraints (and maybe time constraints) are specified. The agent or experiment designer has knowledge of all the givens (including costs of reagents) for the protein that needs to be crystallized. The task of the designer is then to decide how many experiments to perform in parallel, and what the actual values of the controllable parameters are for each experiment.

If this same task needs to be repeated in light of partial outcomes that have been observed and evaluated, the type of decision-making involved changes slightly. The designer now needs to make a choice among several possible actions, basically involving whether to begin a new trial or not. These include:

1. Do not start a new trial, but observe *some* or *all* existing experiments after a certain period of time, when new evaluation can take place.
2. Close *none*, *some* or *all* of the existing experiments for observation and start a new trial, which requires further decision-making as to how many experiments to setup and what controllable parameter values to use for each.

The decision-making process is thus complex and involves tradeoffs along several dimensions, two of them being cost and time. If the decision made is to design a new trial, then the user must enter information regarding the number of experiments to perform in parallel, the number of values for each control parameter that she would like to vary across the experiments, and specify the values. In this prototype implementation of the PEP framework, we include only one type of search or manner in which the specified sets of values for each control parameter are spread over the set of parallel experiments. This is essentially based on a factorial grid where each possible combination of values that could be generated from the specified sets of values for control parameters is used to represent a single experiment. Thus, the grid or factorial search generates a set of parallel experiments that comprise a single trial.

There are several types of initial search methods that could be included as part of the framework in the future. These represent different types of initial screen designs from prior research. A well known incomplete factorial approach (Carter & Carter 1979) assumed that all points in the parameter space of crystallization conditions are equally probable, which has been proven to be an incorrect assumption (Hennessy *et al.* 1999). Jancarik and Kim (Jancarik & Kim 1991) employed a semi-automated sparse matrix sampling of published crystallization conditions that has led to the design of commercially available crystallization kits that contain several initial experiments that have proven to be fairly successful. Grid screens (Weber 1991) are factorial designs and constitute a popular method for systematic screening of crystallization conditions.

**Generate Partial Result** Generation of partial results refers to the task of observing and reporting (electronically) the partial result observed for each experiment. An assumption is made herein that since the process is slow, the exact time at which an experiment is observed could be slightly different than the time noted, since a whole bunch of experiments are typically observed sequentially by a single agent. (If there were multiple agents, then true parallel effects can be achieved.) If we want to get very sophisticated in terms of modeling time, then we could increment current time after each observation, and after each setup. For our purposes, it is sufficient to assume that the same time-lag for setting up each subsequent experiment, balances the time difference in observation of the experiments in a sequential manner. So, we can assume that each trial starts at a particular time, and all experiments within that trial start at that same time, that is simultaneously.

In the laboratory typically, experiments are checked on a daily basis, though sometimes, they can be left unobserved for a week or so, depending on what is known a priori about both the protein being crystallized, as well as the experimental conditions themselves. The types of partial results that are observed can be categorized broadly into one of three types - clear (i.e., no change), crystal, or precipitate. Precipitates and crystals have further categorizations.

We have chosen the parameters of the simulation model such as to produce quantitative output describing partial result quality that resembles the actual resolution limits of diffraction (difflims) used for describing crystal quality in the laboratory. Precipitates are described as large numbers. Good crystals have difflims less than 3Å.

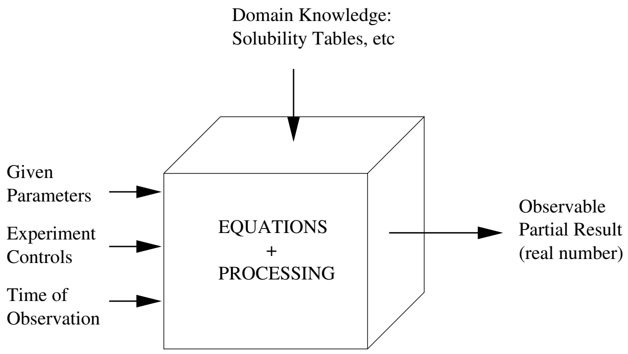


Figure 8: Overview of simulator model

This model is a major component for providing intelligent assistance and hence has been described in detail in another paper (Gopalakrishnan, Buchanan, & Rosenberg 2000). The model contains functional definitions and procedural descriptions of the overall process of protein crystallization. Figure 8 presents an overview of the model used to generate different hypothetical protein crystallization behaviors. By simply varying the controllable parameters that are input to the simulator, it is possible to simulate large numbers of virtual experiments and their outcomes over time, for a particular set of given parameters and domain knowledge. By varying the given parameters and some critical parameters that are included in the description of the equations and processing that drive the simulation, we can produce the effects of different classes of hypothetical protein responses in different crystallization environments. Thus the model offers both flexibility and power with respect to simulating protein crystallization behaviors.

**Evaluate Partial Result and Generate Experiment Summary** The partial result of an experiment is usually observed visually, under a microscope. The result is a description of the observable(s), i.e., whether a crystal or precipitate is seen, and whether it looks like a needle, plate, or is amorphous, and so on. For the purposes of our initial evaluation, we classify the simulated partial result broadly into three classes: clear, precipitate and crystal. It should be pointed out at this time, that the model that describes protein response behavior in different physico-chemical settings is flexible and sophisticated enough to be able to describe and produce different types of solid phases of a protein, each with varying description in terms of shape and size. For purposes of our analysis, we will use just the simple classification, and produce simulations modeling mainly two types of solid phases, i.e., crystalline and amorphous. The amorphous solid phase is labeled as *precipitate*. The crystalline solid phase is labeled as *crystal*. The liquid phase is labeled as *clear*.

1. Update the number of observations performed for experiment E
2. If first observation, compute direct costs (trial T, experiment E)
3. Generate summary of observables at current time for E:
   - Time _of_Observation $\longleftarrow$ current_time - start_time(T)
   - Diffraction Limit $\longleftarrow$ Partial Result
   - Result Class $\longleftarrow$ Assign Class
4. Update material consumed (protein mainly)
5. Update observation costs
6. Update best result statistics (best diflim, time obtained)
7. Calculate experiment utility

Figure 9: Evaluate result and generate experiment summary

The procedure used to evaluate the result and generate a summary of an experiment over time is shown in Figure 9. The data structures that comprise a summary of an experiment can be found in (Gopalakrishnan 1999). The utility of an experiment is calculated as follows:
If the experiment is being observed for the first time, we simply assign a number to utility that is one greater than the difference between the largest result class (i.e. 3) and the observed result class. Otherwise, we add to the current utility of the experiment, a quantity proportional to the difference between the result class of current and previous observation and subtract a quantity proportional to current time. The last quantity is

a penalty for delay in obtaining a good quality result.

Thus, the first time experiment e is observed, the following equation is applied:

$$U_{e,t} = 4.0 - R_{e,t} \tag{1}$$

where $U_{e,t}$ is the utility of experiment e at time t of observation, and $R_{e,t}$ is the result class of experiment e at time t. During subsequent observations, the following equation is used to calculate utility of experiment e:

$$U_{e,t} = U_{e,t-1} + R_{e,t-1} - R_{e,t} - (k \times t) \tag{2}$$

where $k$ is a small constant such as 0.1, and t is current time.

**Generate Trial Summary** Figure 10 shows the simple procedure used to evaluate trial results. Tables showing the data structures used to describe a summary of a trial and a trial definition can be found in (Gopalakrishnan 1999).

1. Calculate the percentage of observables belonging to each class
2. Calculate the amount of material consumed and percentage of initially available protein consumed
3. Compute total costs as sum of direct, intangible and observation costs
4. Assess utility if desired

Figure 10: Evaluate trial results and generate trial summary

**Generate Global Information Summary** The data structures that describe the global summary are shown in (Gopalakrishnan 1999). This in essence constitutes a summary of the global state in terms of parallel actions taken and their observed effects. The global information over time summarizes the feedback from the complex environment from multiple probes over time. Part of the global information includes the output from the C4.5 program that takes as input the description of all experiments so far and outputs a model description or theory consisting of a disjunction of conjunctive rules that is consistent with the data seen so far. The features that represent any experiment and its partial result are encoded as a string of comma separated values for each of its control parameters, as well as time of observation of partial result, and finally the class of the partial result (Class 1 means good quality crystal, class 2 refers to poor quality crystal or precipitate, and class 3 means no observable change was found).

**Agent Architecture** One of the key aspects of this PEP framework is that it includes an agent who is the experiment designer in the loop. At the present time, the agent is a human. We will now explain the characteristics of an agent that resides within the loop in terms of the types of inputs (percepts) that will be received and the types of actions that the agent will perform through its effectors (such as hands that can enter information into a computer from a keyboard, or set up new experiments in the lab).

**Agent Function**

The inputs to the human agent involve the percepts from the environment -

1. Raw observation labels for each experiment at time of observation,
2. Summaries for each experiment, trial, and global state, which contain costs as a sum of setup and observation costs for each experiment, trial and global state information, as well as the amount of resources consumed, and
3. A representation of the evolving boundary in the form of a disjunction of conjunctive rules that describe a model of overall protein response behavior over different times of observation over all experiments performed so far until current observation time.

Assume that the agent has knowledge of its perceptual history and uses some policy[4] for deciding its next action. We also assume that the human agent is able to ascertain whether the goal has been achieved based on the feedback from the environment, that is the percepts.

The outputs or actions that the human agent within the PEP framework performs upon its simulated environment are:

1. Exit (that is, say no to all possible actions) OR
2. Re-open or close experiments or trials for next observation (optional) AND
3. Design a new trial AND/OR observe experiments after some time period.

Figure 11 depicts a commonly used policy by crystallographers for deciding the next set of experiments to perform based on partial result examination. This policy was determined after several long discussions with crystallographers who perform laboratory work on a day-to-day basis. The challenging aspect of the decision-making involves heuristic 3, where the boundary between partial results can help identify regions that have a higher probability of success. The reason

[4] See Figure 11 for a commonly used policy elicited from discussions with several crystallographers.

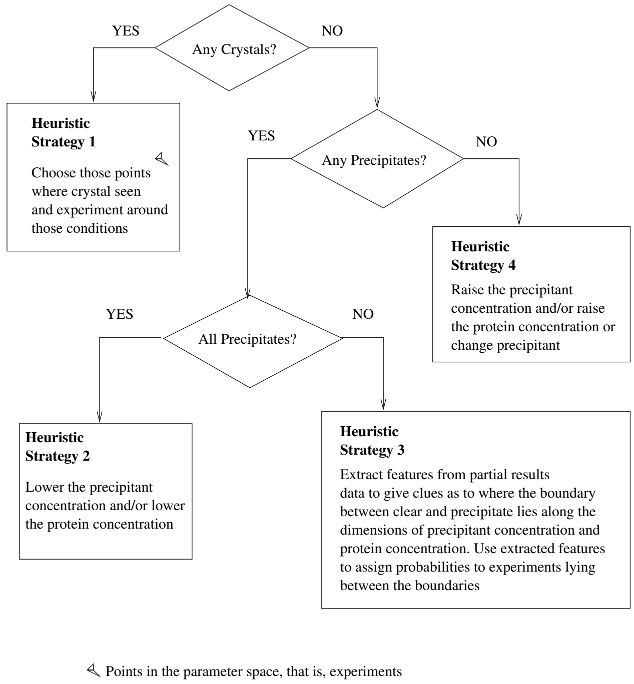


Figure 11: Binary decision tree for experiment design

that boundary identification is very useful in this problem domain is evident from the characteristics of the solution as shown in Figure 1. Intelligent assistance with respect to boundary identification and description is therefore highly desirable. In our experiments with the PEP system using the simulator, we found the policy in Figure 11 to hold for most cases of hypothetical protein givens. This appears to validate both the simulator model as well as the policy, as they were both constructed independent of one another and involved different sets of crystallographers.

## Summary of PEP Framework Implementation

The types of goals for which this framework was designed are:

G1: Identify and label experiments as belonging to particular classes (1, 2 or 3) at time of observation based on partial results.

G2: If crystals are found, identify and report the experimental conditions that seem favorable.

G3: Identify the boundary between the classes along the most influential variables. The most influential variables themselves will need to be identified.

G4: Calculate the costs involved for setup and observations of each experiment and the various trials, and use them for deciding what types of tradeoffs to make at each decision-step.

G5: Keep track of resources that are being consumed, particularly that of protein. This will directly influence strategy.

G6: Store summaries of effects of parallel actions in a data structure (such as the global summary), for analysis of strategies.

Goal 1 (G1) is achieved based on the final quality outcome that is obtained from one run of the simulator. Usually class 1 or good crystals are outcomes of 3.0 or less. Class 2 refers to either bad crystals or precipitates that are outcomes between 3 and 100. Class 3 refers to an outcome that indicates a clear result - 1000.

Goal 3 (G3) reflects the identification of the boundary between the classes of observed outcomes. This goal is achieved using inductive machine learning programs such as C4.5 and RL, that provide a human-understandable description of the boundary based on all the experiments and outcome labels seen so far. Identification of boundary is an important aspect of being able to come up with some internal model of how the given protein seems to be responding under different experimental controls. This gives some insight into possible rates of change of various hidden variables that influence crystal nucleation and growth such as the saturation level. Human agents do not necessarily think about the process in terms of the kind of influence some input controls might have on hidden variables. They focus on the kinds of boundary separations they observe, and based on their knowledge of the process based on experience, they tend to make reasonable hypotheses that influence their search toward more profitable areas.

The methods used for achieving G1 and G3 provide the intelligent assistance for designing crystallization experiments. The remaining goals are met by the PEP information management system.

## Evaluation of the PEP Framework

The central claim is that the PEP framework is sufficient for parallel experiment planning. Sufficiency of this framework is demonstrated by constructing a working program called the PEP system that allows such an evaluation. For purposes of discussion, we will view the framework as consisting of an environment and an agent. The entire framework then supports parallel experiment planning. The design of the environment is based on its utility with respect to the agent that is included in the framework. In the current PEP system, the environment serves as a tool using which a human agent could explore and learn by observing effects of actions. It is immediately obvious that the nature of the feedback from the environment affects the kinds of reasoning tasks that the human agent would need to perform.

The sufficiency of the PEP framework has been shown by constructing an experiment simulator and embedding it within the model that represents the parallel experiment planning environment. This simulated environment interacts with a human agent, the experiment designer, by asking questions. The human agent interacts with the environment by answering the questions and typing in his/her choices for parallel actions. The environment simulates partial results at user-defined observation times and provides the user with experiment summaries, parallel action or trial summaries, and a summary of the state of the world or global summary after each observation action is performed by the user.

The reason that this implementation of the PEP framework works is because:
(a) the main components have been identified and represented adequately,
(b) the instruments for manipulation and interpretation include computational procedures and bookkeeping as well as a human designer,
(c) the environment is controlled and hence, errors such as incorrect recording of observed outcomes do not occur,
(d) the interaction between environment and agent is facilitated smoothly and easily in a question-answer fashion that most humans are used to and seem to like, and
(e) even though the experiment simulator developed in this research is fairly sophisticated, it is possible to find a solution if one exists within a reasonable amount of time expended by a human-agent – the average time taken to learn to use and find one hit for a medium-difficulty simulated protein is about 1.5 hours of real time.

The PEP framework allows for the representation of any number of parallel experiments, that can be represented by data structures that describe:
1. the number of experiments N to be performed in parallel,
2. a set of constant values for all the experiments, that represents the *givens* of a protein or problem,
3. a set of values for control parameters of each experiment, that vary across the N parallel experiments,
4. a search type that represents a program or algorithm used for deciding how to vary the control values across the N experiments,
5. a variable cost function that depends on N that represents the difficulty associated with performing large numbers of experiments in the laboratory, and
6. constraints on material resources, such as amount of protein available. Future resources can also be represented if necessary. Thus, the constraints on resources could change dynamically.

The simulation model is used to generate the partial result observed for each individual experiment. The PEP system with the experiment simulator has been specifically tuned to reflect very closely the way experimentation is actually carried out in the laboratory (that is, the types of commonly used values for many of the major variables in an experiment). A simple evaluation function can convert the quantitative outcome of the simulated partial result into one of three commonly observed classes of results, namely clear (or no result), precipitate, or crystal.

Summaries are provided at all levels of detail, starting with the partial results of an experiment over the observations so far. The most important aspect of the PEP framework is the flexibility with respect to being able to utilize different policies for acting given an intelligent analysis of all partial results observed so far. Thus, it is possible to employ different tradeoffs at any step in the decision path, and observe the outcome of such an action (parallel action) and be able to adjust the weight of possibly performing such an action from the given state. In essence, this PEP framework can be used to effectively learn policy (or policies) that essentially enables the agent (or designer) to decide on the action that is more likely to take him/her closer to the goal. Interestingly, the machine learning component plays a big role in enabling the agent to form a representation of this policy by generating rules that help discriminate effects of actions.

The local policy for choosing next moves changes according to the partial results seen and past experience of the experimenter. It is difficult to automate entirely the analysis of partial results, and hence the need for a human in the loop. The main assumption behind this inclusion is that humans tend to employ different tradeoffs at different stages in the decision making process, which are difficult to enumerate and encode due to the enormous branching factor of related actions and the tacit knowledge of past experience.

## Demonstration of Sufficiency using Prototype

The prototype PEP system has been evaluated by including humans in the loop. The subjects included 2 novice crystallographers, who were introduced to the PEP system for the first time. They were asked to assign a quality number between 1 and 10 for certain evaluation criteria. The summary is shown in table below and clearly indicates that the PEP system is sufficient as an information management tool, and possibly as an intelligent aid to decisions. The subjects found the rules from the machine learning program to be a very simple and easy-to-understand description of the boundary between the classes of observable results. This description was influential in guiding the crystallographers toward the more likely regions for successful crystallization of a particular hypothetical protein. The subjects were also able to detect differences between types of hypothetical protein behaviors, as was visible from the effects produced by the simulator.

| *Question* | *Comments* | *Score (1-10)* |
|---|---|---|
| Interesting? | Yes, makes life easier Summaries are especially helpful at all levels. | 9 |
| Useful? | Very useful because it is easy to use. Easy to follow - clear and concise. | 9 |
| Interface? | Interactive - do not need a graphic - easy to understand already Rules are very helpful - limit your choices and help you get a crystal faster. | 9 |
| Overall Score? | Sufficient for PEP | 9 |

Table 1: Overall evaluation of PEP prototype by the human subjects

## Future Work

The PEP system developed in this paper could be used an educational tool for novice crystallographers. The time complexity analysis of the algorithm that underlies the environment of parallel experiment planning with intelligent decision-making is O(#observations) at each decision step, if we assume a parallel implementation. The current implementation is serial and will need to scale up.

The PEP framework could also be applied to other domains, such as clinical trial design for testing drug efficacy and safety, where the goal is to minimize dosage and maximize benefits while trying to distinguish between toxic and therapeutic effects of a drug.

## References


Bergfors, T. 1990. The crystallization lab manual. Technical report, Biological Sciences, University of Pittsburgh.

Carter, C., and Carter, C. W. 1979. Protein crystallization using incomplete factorial experiments. *Journal of Biological Chemistry* 254:12219–12223.

Ducruix, A., and Geige, R. 1992. *Crystallization of Nucleic Acids and Proteins: a Practical Approach.* Oxford Press.

Feigelson, R. S. 1988. The relevance of small molecule crystal growth theories and techniques to the growth of biological macromolecules. *Journal of Crystal Growth* 90(1).

Gilliland, G. C. 1987. A biological macromolecule crystallization database: a basis for a crystallization strategy. In Geigé, R.; Ducruix, A.; Fontecilla-Camps, J. C.; Feigelson, R. S.; Kern, R.; and McPherson, A., eds., *Crystal Growth of Biological Macromolecules.* North Holland.

Gopalakrishnan, V.; Hennessy, D.; Buchanan, B.; Subramanian, D.; Wilcosz, P. A.; Chandrasekhar, K.; and Rosenberg, J. M. 1994a. Preliminary Tests of Machine Learning Tools for the Analysis of Biological Macromolecular Crystallization Data. Technical Report ISL-94-17, Intelligent Systems Laboratory, University of Pittsburgh.

Gopalakrishnan, V.; Hennessy, D.; Buchanan, B. G.; and Subramanian, D. 1994b. The Crystallographer's Assistant. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1451.

Gopalakrishnan, V.; Buchanan, B. G.; and Rosenberg, J. M. 2000. A model for simulating hypothetical protein crystallization behaviors. Technical report, Intelligent Systems Laboratory, University of Pittsburgh.

Gopalakrishnan, V. 1999. *Parallel Experiment Planning: Macromolecular Crystallization Case Study.* Ph.D. Dissertation, Department of Computer Science, University of Pittsburgh.

Hennessy, D.; Gopalakrishnan, V.; Buchanan, B. G.; Subramanian, D.; and Rosenberg, J. M. 1994. Induction of Rules for Biological Macromolecule Crystallization. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 179–187.

Hennessy, D.; Buchanan, B. G.; Subramanian, D.; Wilkosz, P. A.; and Rosenberg, J. M. 1999. Statistical methods for the objective design of screening procedures for macromolecular crystallization. *submitted to Journal of Applied Crystallography.*

Jancarik, J., and Kim, S. H. 1991. Sparse Matrix Sampling: A Screening Method for Crystallization of Proteins. *Journal of Applied Crystallography* 24:409–411.

Michalski, R., and Stepp, R. 1983. Learning from Observation: Conceptual clustering. In *Machine learning: An artificial intelligence approach.* San Mateo, CA: Morgan Kaufmann.

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

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

Russell, S., and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach.* Prentice Hall.

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

Weber, P. C. 1991. Gridscreen. *Advances in Protein Chemistry* 41:1–36.