

KNOWLEDGE-BASED EXPERIMENT DESIGN IN MOLECULAR GENETICS*

Peter Friedland
Heuristic Programming Project
Computer Science Department
Stanford University
Stanford, CA 94305

A system to produce practical experimental plans in the domain of molecular genetics is in development. It makes use of a large, expert-entered knowledge base containing object-level and strategy knowledge about the domain. The system works by refining general strategies or "skeletal plans," to the level of detail desired by the system user.

1. INTRODUCTION

A large percentage of the experiments carried out in a genetic laboratory fall into the class of structural analysis. The goal of these experiments is to learn something about the physical structure of a particular type of DNA molecule. The experimenter uses his knowledge of fundamental molecular biology, laboratory methods, and basic analytic strategies, combined with what is already known about the structures under study, in order to design an appropriate experiment.

This paper will describe and illustrate one of the methods scientists employ in the experiment design process. It will then discuss a heuristic program which implements this method using a knowledge base entered directly by domain experts.

2. A THEORY OF EXPERIMENT DESIGN

2.1 A Method Scientists Employ

Much of the intellectual effort of experiment design can be characterized as an "encyclopedia-dictionary" loop. Faced with an analytic task, the scientist first picks an appropriate strategy from his collected personal knowledge, from the knowledge of colleagues, or from the literature--his "encyclopedia." The strategy consists of an abstracted or "skeletal" plan which has proven useful for this type of experiment in the past. It may be as general as "label the structural feature you are looking for and then look for the label," or as specific as a detailed laboratory protocol for using an electron microscope to locate loops.

The experimenter then instantiates each of the steps of his skeletal plan with a specific laboratory tool. He does this by first referring to his encyclopedia again, looking for information to help him choose which of the possible techniques is most suitable to the particular problem conditions. Three types of information are useful in making this selection. First, can the technique accomplish the goal of the plan-step? For example, if the specific plan-step were to degrade (break into small pieces) all single-stranded DNA, which techniques are known to have accomplished that task on previous problems. Second, will the technique work on the particular molecule under the particular laboratory conditions? For example, a given enzyme may be able to degrade single-stranded DNA, but it may fail on this particular problem because the molecule is circular or because inhibitors are present in the sample. Finally, which of the remaining techniques is best? This involves using a heuristic based on parameters like convenience, reliability, accuracy, time, and cost.

The refinement process may occur in several hierarchical steps. For the example of degrading all single-stranded material on the molecule, the experimenter might first decide to use enzymatic means of digestion, then decide to use a single-strand specific endonuclease, and finally choose S1 endonuclease. The advantage of this hierarchical approach is that the experimenter does not have to consider all techniques as independent, instance-level entities. Instead of making a choice among, say, one hundred different degradation methods, he can first make use of heuristics to choose between two basic classes, enzymatic and nonenzymatic. After selecting enzymatic, he decides between endonucleases and exonucleases, and then between double-strand-specific and single-strand-specific endonucleases. His final decision is among only half a dozen or so particu-

*This research is supported by NSF
MCS 78-02777.

lar enzymes.

Sometimes the experiment designer will not be able to locate a suitable instantiation of a skeletal plan step. Then he refers to his "dictionary," attempting to understand what is the purpose of the step so that he can break it up into its component parts. For example, suppose the experimenter cannot find a workable instantiation for a denaturation step. He uses his dictionary to find that denaturation really means breaking the hydrogen bonds between the two strands of DNA. He then attempts to find a new skeletal plan for the subgoal of hydrogen bond breaking and repeats the experiment design process for that skeletal plan. The idea is that he is now treating denaturation as the goal of an experiment design problem, i.e. that subgoals are generated when instantiation fails.

2.2 An Example of Plan Instantiation

An example of how a single, general skeletal plan can lead to the design of different experiments can be seen in the problem of sequencing. Sequencing is the process of determining the location on a molecule of specific sites, usually either bases or specific strings of bases. A general strategy for sequencing is as follows:

1. Label one end of the structure.
2. Treat the structure so that on the average one site is cut per molecule.
3. Determine the length of the labeled fragments.

This strategy has been instantiated in different ways for the two problems of total base sequencing and restriction site mapping. For total base sequencing the problem is to determine which of the four bases, A, G, C, or T, is present at each position on the molecule. The first step in the plan is refined to labeling with radioactive Phosphorus 32, the current best method. The second step is instantiated either by a series of chemical reactions or by a carefully controlled enzymatic reaction. The final step is accomplished by separating the fragments by electrophoresis and locating the labels with autoradiography. The process is known as Maxam-Gilbert sequencing (1) in the chemical case and Sanger-Coulson sequencing (2) in the enzymatic case. Both are thought of as major advances, decreasing by over an order of magnitude the time required to determine the base sequence of moderately complex molecule from the previous best method.

For restriction site mapping, the goal is to locate specific four-, five-, or six-base

strings on a structure. (These are the active cutting site for enzymes known as restriction enzymes). The first and third steps of the skeletal plan are instantiated as above; the second step is refined to cutting with a partial digest of the desired restriction enzyme (3).

2.3 Relation to other AI Work

The basic idea of maintaining a collection of abstracted plans relates closely to Schank's concept of scripts in understanding natural language (4). The human scientist, in the typical experiment design case, does not plan from scratch; instead he makes use of a library of knowledge about how classes of laboratory methods fit together to accomplish certain design goals. Maintaining a collection of parameterized plans was first discussed in ABSTRIPS with the name of MACROPS (5). The benefits of hierarchical plan refinement have evolved from ABSTRIPS (6) planning work through the robot planning systems of many other including Sacerdoti (7).

3. IMPLEMENTATION

3.1 The Experiment Design Heuristic

The planning method has been implemented in a straightforward manner. An experimenter describes his problem by telling what he already knows about the structure and its surrounding environment and providing a goal for the experiment. The system searches its library of strategies for skeletal plans that have proven useful for satisfying the given goal or ones like it in the past.

The experiment design system then refines a skeletal plan by working through a detailed taxonomy of laboratory techniques. It uses domain-specific rules to carry out the three-fold selection process described above: goal-oriented rules to choose techniques which can accomplish the desired task, rules relating technique parameters to molecular and sample conditions to eliminate those techniques which will not "go" for the given problem, and heuristics for making a final selection based on the relative convenience, reliability, accuracy, and cost of the competing techniques.

The system also makes use of rules which point out possible interactions between plan-steps. When it encounters an applicable rule of this sort, it postpones further refinement of the step until enough information is known to minimize an unfavorable interaction or (the more common case) maximize the favorable effects of

an interaction.

If an appropriate technique cannot be found to instantiate any plan-step, the system recurs as described above. A subgoal is created to design an experimental plan to provide a new method for accomplishing the plan step.

3.2 The Knowledge Base

All three of the processes described above--problem description, skeletal plan finding, and plan-step refinement--are heavily dependent upon a detailed knowledge base. This knowledge base is being constructed by molecular biologists and biochemists through the facilities of the Unit Package developed at Stanford (8).

In the problem description phase of the process the user is guided in creating an instance of a unit for a DNA structure. The prototypical DNA structure unit contains slots for all the possible kinds of information that could be known for DNA structures. It also contains rules for checking consistency of information and for filling in additional slots from given ones--users don't always realize how much they really know about their structures. The user then describes other laboratory conditions by creating an instance of a prototypical sample-conditions unit.

The skeletal plans are each described as units in the knowledge base. The units contain a slot listing the possible general and specific utilities of the plan, as well as a step-by-step description of the plan. For example, the skeletal plan for locating inverse complementary regions on all types of structures is:

1. DENATURE
2. RENATURE RAPIDLY
3. DEGRADE SINGLE-STRANDED
4. DETECT LARGE-MOLECULAR-WEIGHT DNA

The bulk of the knowledge base consists of the descriptions of laboratory techniques used by the plan-refinement part of the experiment design system. The technique descriptions are arranged in a tree with the major branches being sequence-analysis, modification, separation and detection. Each class of techniques or individual technique is described in a unit containing declarative information about utility, optimal laboratory conditions, allowable substrates and the like, as well as the selection rules that are relevant to that technique.

The knowledge base was constructed by several Stanford molecular biologists using the interactive facilities of the Unit Package. It

currently contains approximately 25 skeletal plan units and 200 laboratory technique units. There are about 250 rules for plan-step instantiation, consistency checking, and modelling the consequences of modification techniques.

4. RESULTS AND CONCLUSIONS

The experiment design system is currently operational and seems to give satisfactory (i.e., plausible) results for a variety of analysis experiments. The particular experimental problems of finding self-complementary regions (inverted repeats), and mapping intervening sequences between genes have received the most attention. As the knowledge base continues to grow in scope and complexity, the range of experimental plans produced should also increase.

An automated experiment design system will serve two major functions: thoroughness and variety. Some instantiations of experimental strategies are missed either because a scientist is not aware of all relevant techniques or because of prejudices toward certain techniques. The design system will be as thorough as its knowledge base--the product of many experts' efforts. The system will promote variety, as it is able to combine the best strategies of different experimenters through the recursive encyclopedia-dictionary process.

REFERENCES

- (1) Maxam, A. and Gilbert, W., PNAS, 74, 1977, 560-564.
- (2) Sanger, F. and Coulson, A., PNAS, 74, 1977, 5463-5466.
- (3) Smith, W. and Birnstein, M., Nucleic Acids Research, 3, 1976, 2387-2398.
- (4) Schank, R. and Abelson, R., Scripts, Plans, Goals and Understanding, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.
- (5) Fikes, R., Hart, P. and Nilsson, N., "Learning and Executing Generalized Robot Plans," Artificial Intelligence, 3, 1972, 251-288.
- (6) Sacerdoti, E., "Planning in a Hierarchy of Abstraction Spaces," in Proc. IJCAI-73, 412-422.
- (7) Sacerdoti, E., "A Structure for Plans and Behavior," Ph.D. Thesis, SRI Artificial Intelligence Center Technical Note 109, 1975.
- (8) Stefik, M., "An Examination of a Frame-Structured Representation System," HPP Report HPP-78-13, 1978.