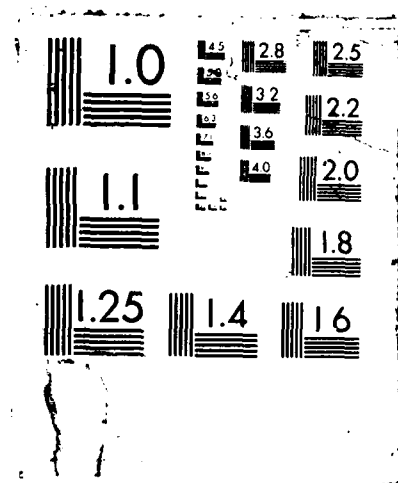ATIC FILE COPY ④

AD-A189 328 — REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	

4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION Knowledge Systems Laboratory Stanford University	6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION Personnel and Training Research Programs
6c ADDRESS (City, State, and ZIP Code) 701 Welch Road, Building C Palo Alto, CA 94304		7b ADDRESS (City, State, and ZIP Code) Office of Naval Research (Code 1142PT) 800 North Quincy Street Arlington, VA 22217-5000
8a NAME OF FUNDING / SPONSORING ORGANIZATION	8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0652

8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO 61153N	PROJECT NO RR04204	TASK NO RR04206-OC	WORK UNIT ACCESSION NO NR442c019

11 TITLE (Include Security Classification)
Final Technical Report, Modeling Expert Control Knowledge   Unclassified

DTIC SELECTED DEC 28 1987

12 PERSONAL AUTHOR(S)
Craig W. Cornelius

13a TYPE OF REPORT Final Technical	13b TIME COVERED FROM 6/4/86 TO 6/3/87	14 DATE OF REPORT (Year, Month, Day) 12/10/87	15 PAGE COUNT 8

16 SUPPLEMENTARY NOTATION

17	COSATI CODES		18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	control reasoning; artificial intelligence, expert systems, blackboard architecture; knowledge-based systems; PROTEAN, protein structure determination
05	08		

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This research explored control knowledge in problem-solving in the domain of protein structure analysis. The objectives were to capture strategic knowledge for this problem, implement this knowledge in working expert systems, and measure its effectiveness.

We designed, implemented, and performed experiments on several control strategies for two aspects of the domain problem. The strategic knowledge was obtained in two ways: by (1) active participation of chemists experts in the domain, and (2) by inferring strategic approaches from published papers and books. We tested our captured control knowledge by measuring its performance under several variations of the strategies to compare the efficiency of problem solving with different amounts and kinds of knowledge. In addition, we compared results from our program with published results using by other means.

Important aspects of this work have been generalized to other problem domains. We find that much of the control knowledge developed from work on protein structures is easily applicable to other constraint satisfaction problems.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified/unlimited
22a NAME OF RESPONSIBLE INDIVIDUAL Dr. Susan Chipman	22b TELEPHONE (Include Area Code) 202-696-4318	22c OFFICE SYMBOL ONR 1142Pt

DD FORM 1473, 84 MAR   83 APR edition may be used until exhausted   SECURITY CLASSIFICATION OF THIS PAGE
All other editions are obsolete

Final Technical Report

Modeling Expert Control Knowledge

ONR Contract N00014-86-K-0652

Knowledge Systems Laboratory
Stanford University

10 December 1987

Accession For	
NTIS GRA&I	☒
DTIC TAB	☒
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

87 12 14 056

Final Technical Report

Modeling Expert Control Knowledge

ONR Contract N00014-86-K-0652

Knowledge Systems Laboratory
Stanford University

10 December 1987

1 Objectives of the Research

Problem-solving expertise comprises two components: expertise in applying particular pieces of domain knowledge to the problem at hand and expertise in determining which pieces of knowledge to apply at each point in the problem-solving process. A growing body of research results characterize expert domain knowledge and problem-solving mechanisms for a variety of problem domains. By contrast, little is known about the knowledge and mechanisms underlying intelligent control.

The research funded under this contract with the Office of Naval Research explored control knowledge in problem-solving in a domain that relies heavily on intelligent choice of actions: protein structure analysis. Our objectives were to capture strategic knowledge for this problem, implement this knowledge in working expert systems, and measure its effectiveness. This work is a part of an ongoing effort to learn general principles of control knowledge and strategy generation and apply them to problem domains with similar characteristics.

2 Summary of Research

Using the "blackboard control architecture" as implemented in the BB1 system[1], we developed and built several control strategies for two aspects of the protein structure problem: finding secondary structure of a protein from experimental nuclear magnetic resonance (NMR) data and determining the placement of major protein structures in three dimensions from empirically determined constraints.

The strategies we encoded were obtained in two ways: by (1) active participation of practicing chemists, experts in determining protein structures, and (2) by inferring strategic approaches from published papers and books in the biochemical literature. We tested our captured control knowledge by measuring its performance under several variations of the strategies to compare the efficiency of problem solving with different amounts and kinds of knowledge. In addition, we compared results from our program that finds secondary structures using several proteins whose structures (determined by other means) had been published in the biochemical literature

In related work, we have begun to generalize our work on protein structures to other problem domains. We find that much of the control knowledge developed from work on protein structures is easily applicable to other constraint satisfaction problems.

3 Research Progress

The follow sections describe the research results from a series of experiments using the

---

[1] Hayes-Roth, B.: "A blackboard architecture for control", *Artificial Intelligence Journal* 26:251-321, 1985.

PROTEAN system, described below. This system is being developed as a collaboration between the Knowledge Systems Laboratory and researchers of the Stanford Magnetic Resonance Laboratory (SMRL) of the School of Medicine, both at Stanford University. Protein structure determination is both knowledge-intensive and computing-intensive, requiring intelligent guidance of the computational process and selection of problem solving techniques to develop solutions that are computable at an affordable cost.

Our research included three main studies to identify, implement, and characterize strategic knowledge in the problem solving process, using the PROTEAN system as the domain problem These included:

- observing the actions of a chemist using PROTEAN and implementing a simple strategy to partially automate experiments in the validation of the PROTEAN system

- building several complete strategies for small proteins, and characterizing each for the costs vs. benefits of control reasoning in the efficiency of finding solutions

- constructing domain and control knowledge for identifying segments of secondary structure in a protein. The expertise encoded in this program ABC was obtained both from interviewing local experts and also from reading papers in the chemical literature.

Below are a short description of the control problem for expert systems, a section describing the PROTEAN system, and a summary of the work in each of the studies mentioned above. We conclude with a listing of relevant technical reports and papers from the Knowledge Systems Laboratory.

3.1 The Control Problem

In attempting to solve a domain problem, a problem solver performs a series of problem-solving actions. Each action is triggered by data or previously generated solution elements, applies some knowledge source in the problem solver's repertoire, and generates or modifies a solution element. At each point in the problem-solving process, several such actions may be possible. The control problem is *which of its potential actions should the problem solver perform at each point in the problem-solving process?* Based upon previous work, in part sponsored by ONR, we have developed a behavior model of expert control problem-solving in which a problem-solver:

- makes explicit control decisions to determine which problem-solving action to perform at each point in the process of solving the problem.

- decides what actions to perform by reconciling independent decisions about actions that should be performed and actions that can be performed.

- uses variable grain-size control heuristics, including global strategies, local objectives, and general scheduling policies,

- adopts heuristics that focus on useful attributes of the current problem-solving situation,

- adopts, retains, and discards individual control heuristics in response to dynamic problem-solving situations,

- decides how to integrate multiple heuristics of varying importance

- dynamically plans, interrupts, resumes, and terminates strategic sequences of actions,

- and reasons about the relative priority of domain and control actions.

We hypothesize that expert problem solvers do not generally apply standard control procedures that anticipate all important situations that may appear in the course of solving the

problem. Rather, they develop control plans incrementally and dynamically, adapting their behavior to a wide range of unanticipated problem-solving situations.

3.2 Constraint Satisfaction and Protein Structures

PROTEAN is a prototype expert system that employs expert knowledge to determine the three-dimensional structure of a protein molecule. Much of the information available on a protein is obtained from nuclear magnetic resonance (NMR) studies, although empirical data from several different biochemical experiments may be used. From NMR, we can find that some particular pairs hydrogen atoms in the protein are positioned within a short distance of each other, thus providing a constraint on the structure of the whole molecule. An NMR experiment can yield several hundred of these distance constraints (known as NOEs, from the Nuclear Overhauser Effect which which these data arise), providing strong limits on the possible shape of the protein.

The PROTEAN system is described in detail in several publications of this laboratory [Lichtarge et al. and Brinkley et al. (KSL 86-28)]. Briefly, the program applies the constraints described above to a protein whose primary structure (i.e. the sequence of amino acids in the protein chain) and secondary structure (sections of local regularity in the structure) are known. The program also applies knowledge that a chemist would apply to elucidate the structure, encoded as knowledge sources in the BB1 system. The result is reported as regions of space in which each part of the molecule can be found, subject to those constraints.

The program is able to work at several levels of detail, first considering large regular structures of the molecule as the basic units (the *solid* level), and then refining a rough solution by considering constraints for individual atoms and bonds of the protein. This two-level approach has significant advantages over methods that immediately work with individual atoms, since solutions at the atomic develop more rapidly when a good approximate structure is used as a starting point.

PROTEAN consists of two main modules: (1) a reasoning system implemented in BB1 that develops a plan and determines what actions are take at each step in the problem-solving, and (2) a geometry system that computes positions of objects consistent with constraints. The geometric computations are time-consuming, even on a powerful computer system, and a bad sequence of actions can result in expensive computations. Since there are many possible ways in which the constraints can be applied, it is necessary to use good strategies for choosing which of the many the possible actions will be most effective in constructing the protein.

3.3 PROTEAN: Control Strategies and Validation

In the work described by Lichtarge et al., we worked with chemists to develop and encode a first test of strategic knowledge for medium-sized proteins. This paper, prepared for the biochemical literature, describes domain knowledge embodied in the system and outlines a simple strategy for applying that knowledge. This knowledge is expressed as control knowledge sources (small, independent program units in the BB1 system) that dynamically plan a strategy that adapts to the kind of protein under consideration. This first strategy was not sufficient to solve the entire protein, but did automatically choose actions for some phases of the solution of the protein. However, more sophisticated parts of the problem-solving required the chemist to specify which of several possible actions was the best to assemble the protein efficiently. The resulting control decisions gave the user the ability to run PROTEAN system in long computations with only occasional guidance.

We performed experiments to validate PROTEAN, i.e. to proves its correctness and determine the behavior of the system at the solid level of representation. These tested the simple strategy we first developed and also measured the performance of the system by systematically varying parameters of the computation. The results of our tests indicated that PROTEAN builds solutions that accurately reflect the overall three-dimensional structure of a protein. The solutions found depend on the quality of the input data, but are comparable to those obtained by other methods of structure determination. The strategic decisions actions of PROTEAN in

this study were partly automated, requiring the user to rate some kinds of assembly actions, while others were suggested by the system. The control strategy developed in this validation effort was used as the basis in the next phase of this work, described in the following section.

3.4 Costs and Benefits of Control Knowledge

The construction of a protein with PROTEAN requires two components: (1) a geometry system (GS) that determines positions of objects subject to constraints, and (2) a program to select a desirable positioning action, initiate a computation in the GS, and interpret the results of these computations. Both components of this PROTEAN system involve computations: (1) the GS searches for legal positions and apply constraints between portions of the structure, and (2) the control system applies heuristics to compute the relative merit of competing feasible actions at each step of the processing. In general, more sophisticated control requires more computation to evaluate and compare the merit of each possible action.

In this study, we developed knowledge sources for PROTEAN that direct the assembly of several small proteins under automatic control. The purpose of this work was to characterize the kinds of control knowledge that were most effective in solving the problem and also to compare the relative costs of control and GS in terms of processing time.

The detailed results of this work are given in Garvey et al. (KSL Report 87-11.) We first developed a basic strategy that completely automates the construction of a solid-level solution for small proteins. This is a dynamic strategy, building a control plan that depends on the characteristics of the problem and that evolves as the problem solving advances. It determines what objects are incorporated in what order in the growing solution and directs what GS routines are called and in what order to position the components of the protein. Although the speed of the construction is critically dependent on the ordering of the actions, the final result of applying the constraints in the problem is *independent* of the detailed ordering of the GS operations.

We studies four variations of the basic control plan that used several different heuristics based on strength of constraints and characteristics of the objects to be positioned. These strategies were applied to one small protein (51 amino acids) and a medium-size molecule (142 amino acids). These additional heuristics were recommended by domain experts to favor positioning actions that rapidly reduce the number of positions for each object, and hence reduce the cost of the computations. However, the rating of the decisions requires significant processing. The main question was: does the benefit of performing fewer but more effective position actions outweigh the cost of identifying those actions?

We found that, indeed, adding more sophisticated heuristics to rate possible positioning actions did indeed increase the speed of the GS operations, particularly when with larger proteins, and also when the GS was directed to computed highly resolved positions for each component. We found that considering the strength of constraints was more effective than using the characteristics of the components when picking a good ordering of the GS actions.

In general, adding control knowledge reduced the total computation time by reducing the number of GS and reasoning operations performed, and also by choosing the GS actions that approach a solution faster. However, in some cases, the added cost of evaluating more heuristics outweighs the added efficiency of computation, particularly when the problem is simple and there is a choice of relatively few actions at most stages of the problem-solving.

3.5 Expert Strategies in ABC

The PROTEAN system described above requires prior identification of regions of secondary structure, since these regions are used as solid-level components of the structure to be determined. ABC is an expert system developed by John Brugge in cooperation with members of the Stanford Magnetic Resonance Laboratory (SMRL). It is implemented in BB1 and offer a clean separation of domain and control knowledge in the interpretation of experimental information.

The secondary structure of a protein can be inferred from data obtained from NMR experiments. Since protein analysis using NMR is a relatively new field, many of the the techniques for interpreting data are heuristic rather than formulaic and often contain biases and assumptions, sometimes implicitly. A knowledge-based system to automating the identification of secondary structures offers the opportunity to explore the representation and use of reasoning processes, as well as to handle the biases and assumptions in a declarative and inspectable form.

The principal method for identifying secondary structure from NMR relies on visual inspection of the data to find meaningful patterns. This technique, known as the *visual method*, does not require complex mathematical calculations but rather depends on the existence of characteristic patterns in the NMR data from the protein. The method was only recently (in the last three years) described completely in the chemical literature, and about one dozen published structures have been determined using this method. It is an informal, heuristic method, using ideas observed to be generally valid but not always correct. The application of artificial intelligence techniques to this problem offers the chance to not only duplicate the expertise of chemists, but to also provide a framework for testing and formalizing new ideas in this field.

In this work, two domain experts from SMRL provided the initial domain and strategy knowledge to begin building a system. Further background reading provided additional heuristics and techniques for the construction of the ABC program, as well as examples and test cases on which the system could be validated.

ABC is implemented in the BB1 system ,and incorporates a dynamic strategy that recommends different actions, depending on the characteristics of the molecule being analyzed. The strategy is explicitly represented on the control blackboard of the system and allows procedural, opportunistic, and heuristic behavior.

The ABC system was tested using data on nine different proteins, six from the chemical literature, and three from known crystal structures. The system identifies secondary structures with an accuracy similar to that of human experts in the field, who themselves do not always agree on the interpretation of data. This was illustrated vividly by three similar but not identical solutions found in two separate published reports and an analysis done by a local domain expert, all on the same protein.

4 Related Research

Related work in the Knowledge Systems Laboratory (KSL) at Stanford University includes SIGHTPLAN, a system for laying out construction sites for building projects. This problem, like PROTEAN, can be framed as a geometric constraint satisfaction problem, in which the objects to be positioned include trucks, building materials, roads, temporary buildings, cranes, and other construction equipment. We have developed an initial prototype of SIGHTPLAN by starting with the control and domain knowledge from the PROTEAN system. Although the constraints and geometric computations in the two domains are far different, we were able to quickly adapt knowledge sources using the ACCORD framework [KSL Report 86-38] to operate in this new domain.

The PROTEAN project is now developing routines to construct a detailed atomic-level solution for small proteins. We are also designing more sophisticated control strategies for larger proteins. In such molecules, independent arrangements of portions of the molecule must first be solved, and then the individual arrangements must be combined into a solution for the entire molecule.

5 Software Products

We have developed two software systems under partial funding from this research contract. PROTEAN is an expert system for determining the three-dimensional structure of protein molecules from empirically determined constraints. Work described in Lichtarge *et al.*

describes software that implements domain and strategic knowledge to model the expertise of a chemist in this task. The prototype system PROTEAN runs as an application within the BB1 system.

ABC is an expert system that captures and expert's knowledge of a related task: identifying secondary structures of a protein from experimental data. ABC is also implemented in BB1 and is described in KSL Report 87-62, by Brugge and Buchanan.

The BB1 system used in both PROTEAN and ABC was designed by Barbara Hayes-Roth. The system is available in both InterLISP and CommonLISP and runs on a variety of computer systems. This system is described in two versions of the BB1 Manual (KSL Reports 86-60 and 86-61).

6 Publications

The following technical reports and published papers from the Knowledge Systems Laboratory include detailed technical descriptions of the research and results summarized in this final report.

1. Olivier Lichtarge, Craig W. Cornelius, Bruce G. Buchanan, and Oleg Jardetzky; **Validation of the First Step of the Heuristic Refinement Method for the Derivation of Solution Structures of Proteins from NMR Data.** Knowledge Systems Laboratory, Computer Science Department, Stanford University, and Stanford Nuclear Magnetic Resonance Laboratory, School of Medicine, Stanford University. July 1987. Accepted for publication in *Proteins: Structure, Function, and Genetics.* 25 pages.

2. KSL 86-28
James Brinkley, Craig Cornelius, Russ Altman, Barbara Hayes-Roth, Olivier Lichtarge, Bruce Duncan, Bruce Buchanan, Oleg Jardetzky; **Application of Constraint Satisfaction Techniques to the Determination of Protein Tertiary Structure,** March 1986. **14 pages**

3. KSL 86-51
Barbara Hayes-Roth, Bruce Buchanan, Olivier Lichtarge, Micheal Hewett, Russ Altman, James Brinkley, Craig Cornelius, Bruce Duncan, and Oleg Jardetzky; **PROTEAN: Deriving protein structure from constraints,** March 1986. Also appears in *Proceedings of the Fifth National Conference on Artificial Intelligence. Philadephia, Pennsylvania, August 1986.* American Association of Artificial Intelligence. Morgan Kaufmann, Publishers. Los Altos, CA. **21 pages**

4. KSL 86-60
Alan Garvey, Micheal Hewett, M. Vaughan Johnson, Robert Schulman, Barbara Hayes-Roth; **BB1 User Manual - Interlisp Version,** October 1986. **68 pages**

5. KSL 86-61
Alan Garvey, Micheal Hewett, M. Vaughan Johnson, Robert Schulman, Barbara Hayes-Roth; **BB1 User Manual - Common Lisp Version,** October 1986. **72 pages**

6. KSL 86-76
M. Vaughan Johnson Jr. and Barbara Hayes-Roth; **Integrating Diverse Reasoning Methods in the BB1 Blackboard Control Architecture,** December 1986. **17 pages**

7. KSL 87-05 and STAN-CS-87-1142
James F. Brinkley, Bruce G. Buchanan, Russ B. Altman, Bruce S. Duncan, Craig W. Cornelius; **A Heuristic Refinement Method for Spatial Constraint Satisfaction Problems,** January 1987. **15 pages**

8. KSL 87-10
Michael Hewett and Barbara Hayes-Roth; **The BB1 Architecture: A Software Engineering View,** February 1987. **17 pages**

9. KSL 87-11

Alan Garvey, Craig Cornelius, and Barbara Hayes-Roth; **Computational Costs versus Benefits of Control Reasoning,** June 1987. Also appears in *Proceedings of the Sixth National* Conference on Artificial Intelligence. Seattle, Washington, July 1987.] American Association of Artificial Intelligence. Morgan Kaufmann, Publishers. Los Altos, CA. **6 pages.**

10. **KSL 87-40**
    Alan Garvey and Barbara Hayes-Roth; **Implementing Diverse Forms of Control Knowledge in Multiple Control Architectures,** June 1987. **32 pages**

11. **KSL 87-62**
    John A. Brugge and Bruce G. Buchanan; **Evolution of a Knowledge-Based System for Determining Structural Components of Proteins,** October 1987. **26 pages**

Distribution List

Final Technical Report

Modeling Expert Control Knowledge

ONR Contract N00014-86-K-0652

Knowledge Systems Laboratory
Stanford University

10 December 1987

The final technical report of this research has been distributed to:

Dr. Susan Chipman   1 copy
Administrative Contracting Officer
Cognitive Science
Office of Naval Research
Department of the Navy
800 North Quincy Street, Code 1142PT
Arlington, VA   22217-5000

Director   1 copy
Naval Research Laboratory, ATTN: Code 2627
Washington, D.C.   20375

Defense Technical Information Center   12 copies
Building 5
Cameron Station
Alexandria, VA   22314

END

DATE
FILMED

MARCH
1988

DTIC