Report 79-16
Stanford -- KSL

Structure and Function of the
CRYSALIS System.
Robert S. Engelmore, Allan Terry,
Jul 1979

STANFORD HEURISTIC PROGRAMMING PROJECT
HPP-79-16

JULY 1979

STRUCTURE AND FUNCTION OF THE CRYSALIS SYSTEM

BY

ROBERT S. ENGELMORE
COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

ALLAN TERRY
DEPT. OF INFORMATION AND COMPUTER SCIENCE
UNIVERSITY OF IRVINE, CALIFORNIA
IRVINE, CALIFORNIA 92717

HEURISTIC PROGRAMMING PROJECT
COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

STRUCTURE AND FUNCTION OF THE CRYSALIS SYSTEM

Robert Engelmore[1]
Computer Science Department
Stanford University
Stanford, California 94305

Allan Terry[2]
Dept. of Information and Computer Science
University of California, Irvine
Irvine, California 92717

ABSTRACT

CRYSALIS is a knowledge-based system whose goal is to infer the three-dimensional structures of proteins from x-ray crystallographic data. The system uses both formal and judgmental knowledge from experts to select appropriate procedures and to constrain the space of plausible protein structures. The hypothesis generating and testing procedures operate upon a variety of representations of the data, and work with several different descriptions of the structure being inferred. This paper focuses on the architecture of the system, and points out some of its interesting features. An example of the system's performance is given.

---

[1] Present address is DARPA/IPTO, 1400 Wilson Blvd., Arlington, Va. 22209

[2] Present address is Dept. of Computer Science, Stanford University, Stanford Ca 94305

Table of Contents

1 Introduction

In this paper we present an application of Artificial Intelligence methodology to the domain of Protein Crystallography.[1] The long term practical goal of this work is the creation of a fully automated system for protein structure determination, starting with data collection and ending with an accurate 3-D model of the molecule. Most of the software already exists at the two ends of the process. At the "front end", programs exist for reducing and transforming the x-ray diffraction data, and estimating phases. The result of this processing is an electron density map (EDM), which gives a blurred view of the electron cloud surrounding the molecule. At the "back end" are a variety of numerical techniques for taking a full or partial model of the protein and iteratively refining the atomic coordinates so that the model is a best compromise between one which best fits the data and one which best matches ideal stereochemical constraints [Hermans74] [Agarwal77]. The "middle" portion of the process is the generation of a first-order model of the protein, based on an interpretation of the EDM and the amino acid sequence (when known). EDM interpretation has traditionally been a manual process of visual pattern recognition, for which CRT displays have become a significant aid in recent years. Although the use of a large computing system in conjunction with a graphical display has made a significant reduction in EDM interpretation time [Tsernoglou77], the knowledge used in matching the molecule to the EDM remains in the head of the model builder. Our aim is to capture that knowledge within an automated EDM interpretation program, thereby closing the gap which remains in building a fully automated system.

Many diverse sources of knowledge contribute to the inference of a protein structure from an EDM and associated data. Examples: chemical knowledge about protein composition is used to estimate overall size and to generate expectations about special features that should be present in the data; stereochemical knowledge about protein conformation is used to constrain the relative positions of atoms (e.g. certain subsets of atoms form rigid groups); X-ray crystallographic knowledge is used to detect symmetries, or to match a hypothesized structure against the data; specific heuristics for interpreting EDMs are continually invoked to match features in the EDM with expectations from the model.

In order to use the knowledge sources efficiently, a global data base -- the "blackboard" -- is constructed which contains the currently active hypothesis elements, at all levels of description. The decision to activate a particular knowledge source is not pre-established, but depends on the current state of the solution and what available knowledge source is most likely to make further progress. The control is, to a large extent, determined by what has just been learned: a small change in the state of the "blackboard" may establish a new island of opportunity, providing the preconditions to instantiate further knowledge sources.

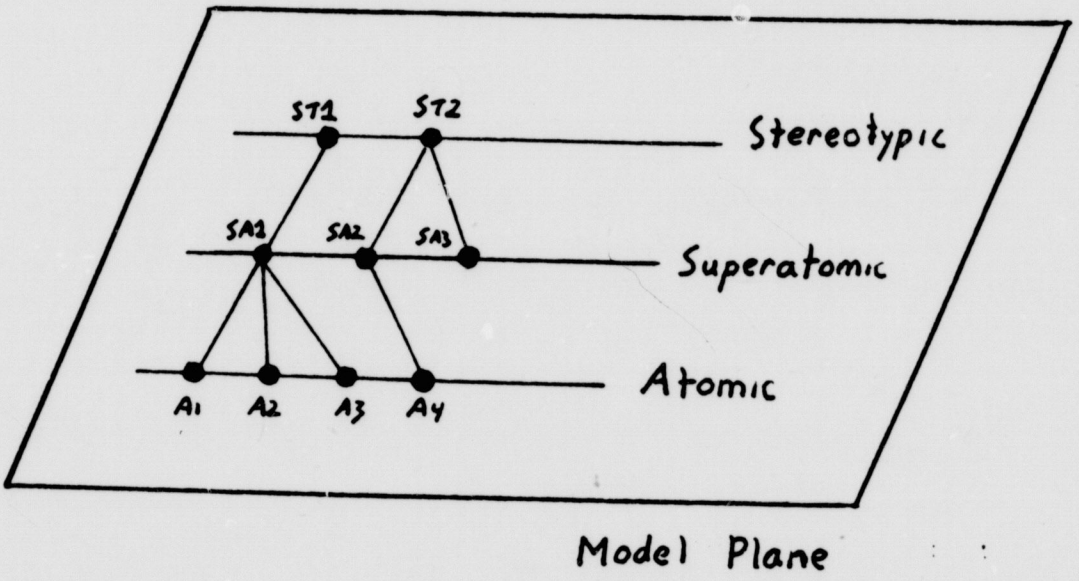
Figure 1 shows the types of data and hypotheses that are used in CRYSALIS. As in Hearsay-II [Erman76], the hypotheses are represented in a hierarchically organized data structure. In our case the different information levels can be partitioned into two distinctly different

---

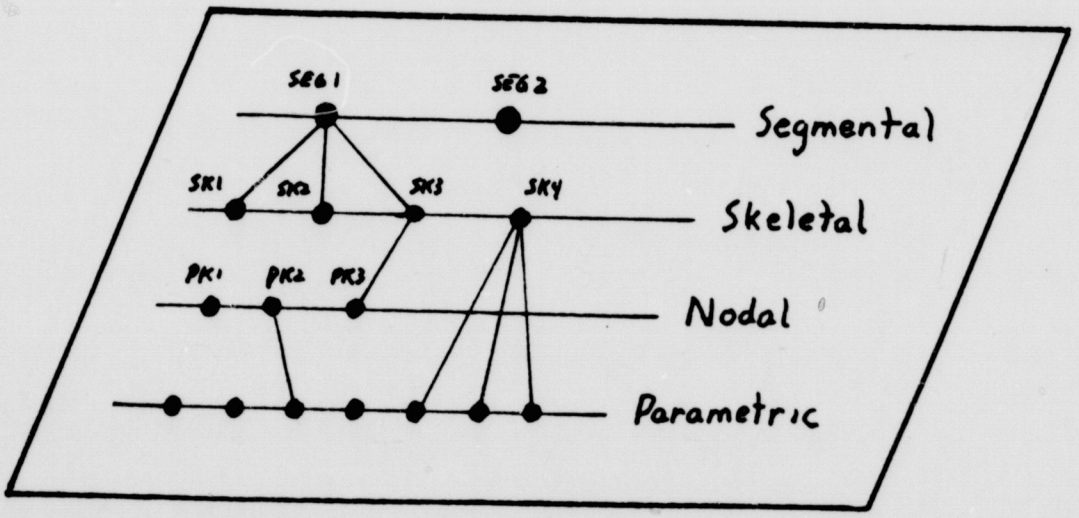
[1] This work was supported by the National Science Foundation (MCS74-23461-A01) and the Advanced Research Projects Agency (MDA 903-77-C-0322). Use of the SUMEX computer facility was made possible by the Biotechnology Resources Program of the NIH under Grant RR-00785.

"planes", but the concept of a globally accessible space of hypotheses and data abstractions is essentially the same for both systems. Knowledge sources play the same role as in Hearsay-II, adding, changing, or testing hypothesis elements on the blackboard. For further discussion see [Engelmore77]. The processes of generating or modifying hypotheses and of invoking knowledge sources are nearly identical to those described for the AGE system [Nii79].

Fig. 1



Model Plane



Density Plane

In the course of deriving a protein structure which is a best explanation of the given data, the crystallographer generates a three-dimensional description of the electron density distribution of the molecule, often called an electron density map (EDM). Due to the resolution imposed by the experimental conditions, the EDM is an indistinct image of the structure, which does not reveal the positions of individual atoms. The crystallographer must interpret the map in light of auxiliary data and general principles of protein chemistry in order to derive a complete description of the molecular structure.

The most important piece of auxiliary data is the amino acid sequence, along with a specification of any special chemical groups attached to the protein chain. This information defines the set of atoms which comprise the structure, and their interconnections, but gives little or no information on how the protein folds and twists from one end of its polypeptide chain to the other. The EDM, on the other hand, shows where the protein must lie in 3-space, but gives little information on which atoms go where. Thus the amino acid sequence and and the EDM are complementary sources of information.

By carefully studying the map, and guided by the amino acid sequence, the experienced protein crystallographer can find features which allow him to infer approximate atomic locations, molecular boundaries, groups of atoms, the polypeptide backbone, etc. Typically, several weeks to months of tedious effort are required to build, manually, a model of the molecular structure which conforms to the electron density map and is also consistent with the sequence, his knowledge of protein chemistry, stereochemical constraints and other available chemical and physical data.

2 Structure of the CRYSALIS system

In the CRYSALIS system the data, the hypothesis and the knowledge base are all hierarchically structured. The data have several levels of abstraction, the hypothesis is described at three levels of detail, and the knowledge base is divided into domain knowledge, task knowledge and strategy knowledge. These three hierarchies are discussed below.

2.1 The data hierarchy

The Electron Density Map

The primary input data source is the EDM, which is derived from X-ray diffraction studies of the crystallized protein. The EDM is an image of the electron cloud surrounding the atoms of the molecule. Its representation is a set of intensity values defined on a three-dimensional grid of 100,000 to 300,000 points. Usually 1% to 6% of the intensities are significant to a model builder. The usefulness of the map can be characterized by its resolution and its quality. CRYSALIS is intended to interpret EDMs of resolution in the range of 2.0 to 2.5 Ang. (for comparison, an average distance between atoms is 1.7 Ang.), and of relatively high signal to noise ratio.

Because of the size of the data base, and the relatively small number of significant points, the system uses several abstractions of these data.

The Peak List

The most obvious abstraction of the map is a list of local maxima. These maxima (peaks) are calculated from the map by a process of interpolation, and are thus not constrained to lie on the EDM grid points. A peak usually indicates the position of an atom, or an average position of a small group of atoms. The height of a peak is a rough linear function of the atomic number of the atom(s) producing the peak. While the inability to resolve individual atoms means there is an error in this correspondence, one can distinguish between various classes of peaks.

The Skeleton

The second major abstraction is the skeleton. By a method first developed by Greer [Greer74] and later refined by us, the EDM is repeatedly scanned, discarding intensities at all grid points which are below a given threshold and not required for preserving continuity of the electron density cloud. What finally remains is a list of grid point locations, called nodes, and their connections to other nodes. In this way one reduces the number of data points by two to three orders of magnitude. The skeleton preserves the general topology of the protein, at the price of losing some of the fine detail that distinguishes one amino acid side chain from another.

The Segment List

The segment list condenses the skeleton (without losing any information) by hiding nodes with only two neighbors. A segment is the set of connected nodes starting at a tip (one neighbor) or branch point (three or more neighbors) and terminating at another tip or branch point.

Once the data have been condensed to the level of segments, one can frequently identify topological features that correspond to structural elements of the protein molecule. For example, there are usually some long, connected segments that correspond to the backbone of the protein. This backbone is sometimes "clean" in that it shows a simple alteration of objects that appear to be sidechains and main chain (peptide) links. Skeletonization is a heuristic procedure, however, and often at least some portions of the skeleton are difficult to interpret.

Subgraphs of the Segment List

It is often useful to think in terms of subgraphs of segments. Since in many cases these subgraphs correspond to superatoms (see below), many of the heuristics for EDM interpretation can be expressed in terms of subgraphs. There are two major kinds of subgraphs: sidechains and peptides. Sidechain subgraphs are collections of segments which look like they might be representations of real sidechains (a bit of nomenclature: a "sidechain" is the physical entity the system is trying to model, a "sidechain subgraph" is a part of the segment list that one suspects is the skeletal representation of the sidechain).

Thus far we have identified six kinds of sidechain subgraphs. Peptide subgraphs are defined as all segments between one sidechain subgraph (or bridge) and the next.

2.2 The hypothesis hierarchy

The goal hypothesis in our system is a model of a protein molecule which best explains the given experimental data and is consistent with accepted principles of stereochemistry and protein chemistry. As mentioned earlier, many diverse sources of knowledge are brought to bear on the problem of EDM interpretation. Each knowledge source (KS) may use different descriptions of the objects on which it operates. For example, a helix locater works with an abstraction of a molecule consisting of a specification of the backbone shape, omitting all other details. A side chain template matcher, on the other hand, uses a full specification of all atomic positions to define the side chain template. In order to cope with this diversity, the hypothesis is represented as hierarchically organized levels of descriptions, as shown in Figure 1. A KS is a collection of rules which makes inferences either within or between levels in the hypothesis space.

There are three levels of description in the model plane. The most detailed level is the atomic level, a specification of the spatial coordinates of all atoms in the model with respect to some arbitrary origin (the coordinates of hydrogen atoms are generally omitted). Proteins exhibit well-defined topological constraints which permit descriptions at higher levels of aggregation. Thus, proteins consist of a linear polymeric chain and, in many cases, attached atomic groups called co-factors. The level of description which describes the model in terms of familiar groups of atoms such as peptide links, side chains, and cofactors, is called the superatomic level. These units may be aggregated still further into what is generally called a "secondary structure", i.e., a specification of the relative locations of large identifiable portions of the protein. Examples are the alpha helix and the beta sheet conformations, well known to protein chemists. Many other such "stereotypes" exist, although they may be associated with a specific family of proteins. This level of description is labelled stereotypic in Figure 1.

A partial or complete hypothesis consists of linked hypothesis elements. A hypothesis element (HE) is a labelled entity in the space of hypotheses. Attached to each entity is a set of attributes which define the HE in terms appropriate to the level of description on which it resides. HEs at the three levels are defined as follows:
Stereotype Hypothesis Element

LABEL - STn, n = 1, 2, ...
TYPE - e.g. BACKBONE
RANGE - (a . b) where a and b are sequence numbers, and a < b
BLOCK - (x . y); x (or y) is non-NIL if the backbone is blocked at a (or b). A backbone is blocked at an end when the trace that created it is blocked by something in the skeleton it can't interpret.
MEMBERS - a list of superatoms that make up the backbone

CF - a certainty factor equal to the average of all the MEMBER's CFs

Superatom Hypothesis Element

LABEL - SAn, n = 1, 2, ...
TYPE - e.g. PEPTIDE or SIDECHAIN
BELONGSTO - (acid . seq), example is (MET . 3)
MEMBEROF - points to BACKBONE ST, if any
MEMBERS - a list of any atom hypotheses that are part of this superatom
SEGS - list of subgraphs of the segment list and their associated CFs

Atomic Hypothesis Element

LABEL - An, n = 1, 2, ...
TYPE - one of C, N, O, S, Ca, Mg, Fe, CO, Zn, H
NAME - standard name designating position in the residue such as: CA, SD, NE1
BELONGSTO - as in PEPTIDE
MEMBEROF - list of superatoms this atom is in
D.PEAKS - links to peaklist with associated CFs
NODES - links to skeleton with associated CFs
SPACE.LOC - list of possible positions for the atom and their CFs

Note that there may be multiple data links. For instance, an atom may have links to several peaks in the peak list on the data plane, indicating that its position has not been established uniquely. A measure of certainty is associated with each link as it is generated. These certainty factors are used to compare and merge alternate hypotheses.

2.3 The rule hierarchy

The formal and informal procedures which comprise the knowledge sources are expressed as rules. These rules are collected into sets of rules, each set being appropriate to use when particular classes of events occur. The correspondence between event classes and rule sets is established by another set of rules, the task rules. The task rules are used to decide which KS or sequence of KSs to call in order to perform one of the typical tasks in building the structure, e.g., tracing the protein backbone between two anchor points. The decision is based on the state of the blackboard and the items on an event list. The task rules thus form a second layer of rules which direct the system's selection of an appropriate method for proceeding.

Once a task is completed, or if the task fails, the system must look to a higher level of control to determine what to do next. At this higher level -- the strategy level -- the structure building process can either try to solve the current subproblem by another method,

or shift the focus of attention to another region of the structure. Strategy knowledge is expressed as rules which make use of the current state of the blackboard and the event list.

We thus have a completely rule-based control structure, employing three distinct levels of rules (or knowledge): the specialists, commonly called the knowledge sources, the task rules for method selection, and the strategy rules for controlling the focus of attention. Although this pyramidal structure of rules and meta-rules could continue indefinitely, a three-tiered system of control appears to offer sufficient flexibility in choosing and deploying the resident knowledge sources.

The following subsections discuss in greater detail the structure and content of CRYSALIS at the three levels of control.

The strategy level The overall control of the system is assigned to the strategy level. A set of strategy rules governs the choice of task to perform and a region of the hypothesis in which to work. Examples are:

1) IF there are no hypothesis elements
THEN do INITIALIZATION task

2) IF there are two toeholds A and B
and A has certainty greater than 400
and B has certainty greater than 400
and the number of residues in the sequence
separating the toeholds is less than five
THEN do TWO.POINT.TRACE task

(Notes: Certainty is a measure of belief associated with the properties of each hypothesis element. Its range of values is -1000 to 1000. Certainty factors are established and/or modified by KSs as they find support for a hypothesis element in the underlying data. A "toehold" is a hypothesis element that is linked to both the amino acid sequence and to the density plane with a relatively high certainty, i.e. an "identified" part of the EDM.)

3) IF there is a toehold, A
and A has certainty greater than 400
and the direction of the backbone at that toehold
is known
THEN do EXPANSION.TRACE task

The task level At the task level decisions are made on how to best accomplish a specific task, i.e. which KS or combination of KSs should be used. This knowledge is embodied in a set of task rules, whose conditions are predicates operating on the event list, and whose

actions invoke specific KSs. Eight tasks have been formulated thus far, of which the following are examples:

INITIALIZATION The invocation of this task reads input data into system, puts any given information on the blackboard, tries to identify all large atoms, finds disulphide links, predicts occurrences of helical regions, tries to discover a set of chainends, and performs an "interesting sequence" analysis. If the protein contains cofactors, they are located and removed from the segment list in order to simplify the data. The end result of performing this task is the establishment of a few toeholds in the data from which further inferences can be made.

EXTENSION.TRACE This task is applicable if the system has established the direction of the backbone at a given toehold. When this task is invoked, the model is extended, at the superatomic level, in a given direction until the cumulative certainty drops below a given threshold.

TWO.POINT.TRACE Chain tracing between two given toeholds is in many ways the most certain of the various chain tracing tasks because the limits are known. By tracing from each toehold toward the other and comparing the results, very high certainty values can be generated. This task uses many of the same KSs used in the EXTENSION.TRACE task. The main differences are in the task rules, e.g., determining the conditions for stopping the trace at one side and beginning the trace at the other.

MODEL.DRIVEN.TRACE If it can be established that the current location is part of a known secondary structure (such as alpha helix or beta sheet), a model of that structure can be used to predict successive atomic positions. The trace becomes a very highly constrained loop of prediction and verification. Unlike other tracing tasks, the predictive model allows the trace to skip residues that do not appear in the data.

The knowledge source level The most detailed knowledge of the domain is contained at the KS level. Each KS is a "chunk" of formal or informal knowledge used to solve some particular sub-problem in EDM interpretation. Examples are matching a sidechain subgraph to a template, matching peaks in the EDM with large atoms inferred from the amino acid sequence, or simplifying the skeleton by removing the cofactor contribution. KSs are typically represented as a set of productions, but in some cases it has been more efficient to represent the knowledge as a procedure (particularly when heavy numerical computing is involved). About two dozen KSs have been implemented in the CRYSALIS system thus far.

2.4 Event-driven control

The CRYSALIS system uses an event-driven control structure. An event is a description of a change in the hypothesis, e.g., the addition of a new hypothesis element or the establishment of new links between existing hypothesis elements. In this scheme the current state of the hypothesis space determines what to do next. The normal iterative cycle of problem solving uses the event list to trigger knowledge sources, which create or change hypothesis elements and place new events on the event list. Each task is executed until it is explicitly finished or until no further rules are fired, after which control passes up to the strategy level to shift attention to a new task. The particular hypothesis element under investigation, or the particular KS invoked, are determined by the type of event selected from the event list. Thus, under normal conditions, the monitor always has a means for choosing its next move. Items may be selected from the event list according to a specified processing mode, e.g., FIFO, LIFO or some priority scheme for choosing the "best" event.

The system's behavior is "opportunistic" in that it is guided primarily by what was most recently discovered, rather than by a necessity to satisfy sub-goals. The choice of an event-driven control structure is based partly on efficiency in selecting appropriate knowledge sources and partly on conformity with the structure modeling process normally employed by protein crystallographers.

2.5 Focus of Attention Mechanisms

There are two levels of attention focussing, corresponding to the two levels of control in CRYSALIS. At the strategy level, a "coarse" focus of attention is created by assigning a task to a region of the hypothesis space. The conditions on strategy rules refer to global features of the current hypothesis, such as the presence of solved and unsolved regions along the amino acid sequence. A finer degree of focussing is provided at the task level, where task rules focus on specific events, and specific KSs (where the real model building is accomplished) to process those events.

3 Discussion

The sheer volume and variety of knowledge that crystallographers bring to bear on the EDM interpretation problem implies a correspondingly large and varied automated system. CRYSALIS employs many ideas developed over the past decade for representation and utilization of knowledge in expert systems. Although none of these features by itself breaks new ground in AI research, their conjunction in one system makes an interesting case study.

3.1 Flexible architecture

An essential requirement of CRYSALIS is flexibility: the addition of new facts and procedures must be done easily and quickly, as nearly every new protein structure presents new problems and ad hoc knowledge. The modularity of the rules, the multi-level control and the

multiple blackboard planes for hypotheses and data are all incorporated in the system in order to accommodate expansion or modification. Modularity also allows us as system builders to experiment with alternate designs for the components, e.g. using a strictly procedural knowledge source in place of a rule-based one, with no external change in the system.

3.2 Rule-based control of strategy and task levels

The method of interpreting protein EDM's is, at its critical points, opportunistic. Where to start, when to leave one part of the structure and focus upon another, what level of detail to look at, when to stop -- these questions are continually presented to the expert as he builds his structure. The knowledge needed to answer them is almost entirely heuristic, and as subject to change as any other task-specific knowledge. It thus seems natural, and indeed has shown to be practical, that this strategic knowledge, which controls the order in which various tasks are performed, be represented as rules. Moreover, the tasks may be executed in various ways, depending upon what KSs are available and the particular situation encountered, so control at the task level is also facilitated by a rule-based representation.

The use of two levels of control has been found to be an efficient alternative to using one large rule set.

3.3 Event-driven processing

The opportunities which arise in EDM interpretation are dynamic, i.e., they flow not only from the initial data but from the partially built structure. Thinking of the solution process as one of recognition and verification, the recognition phase operates on both the data and the partial hypothesis. As the hypothesis changes, new islands of opportunity arise. By specifically recording these events, the focus-of-attention problem is ameliorated.

3.4 Multi-level hypothesis/data structure

Use of the blackboard concept was suggested by the similarity of this problem with that of other signal understanding tasks, specifically in the need for a common store accessible to many diverse and independent KSs. The representation of both the hypothesis and the data in the uniform way discussed earlier makes it easier to add or modify the various levels. For example, it might be useful at some time to add a "glob" level of description of the EDM, which would capture some visual information, along with a new level of description of the model, such as families of amino acids. The multi-level planes facilitate this kind of modification.

The multiple levels of description are also necessitated by the lack of a common language for describing the objects and operators of the domain, in contrast, say, to the Heuristic DENDRAL program which uses a uniform language of chemical subgraphs for this purpose. The domain knowledge in CRYSALIS is expressed most naturally at several levels, e.g. inferences

can be drawn from peaks in the EDM to atoms in the model, but inferences about a skeletal segment are related to superatoms.

Moreover, the individual levels of detail in the model plane have an intrinsic interest. As is evident from the protein crystallographic literature, there is more of interest to a molecular structure than the coordinates of its constituent atoms. Announcements of new structures usually contain visual representations of the backbone alone, secondary structure, hydrogen bonding, active site geometry, etc. Often the model is incompletely specified at the atomic level, so that descriptions of the structure in some regions must be couched in the language of side chains or other abstractions. The multi-leveled hypothesis structure in CRYSALIS permits this variety of descriptions.

3.5 Ill-structuredness

Two characteristics of the task domain are particularly interesting in that they provide an ill-structuredness which is normally not found in AI applications. The first is the lack of a well defined termination criterion. One can, of course, stop the program when the position of all atoms is known, but this is almost never the case. In practice, one stops when no further progress can be made, as for example when no strategies available to the system are applicable, or if further model building effort yields diminishing returns (the recognition of which is an interesting and thus far unsolved theoretical problem). Moreover, to the extent that one can lay down some rules for terminating the process, those rules are usually dependent on the quality of the data.

A second characteristic is that the detail of the resulting model, i.e. the solution, varies with the quality and resolution of the given data. Different knowledge sources are invoked for different ranges of resolution, and may either draw different inferences or link different levels of the model plane with the data. In any case the detail of the model is adjusted to fit the experimental information, and can degrade gracefully with decreasing resolution. Thus if some data are particularly poor, or not given at all (e.g. an incomplete sequence) it still should be possible to get some results.

4 Conclusion: Current status and activities

As it currently exists, CRYSALIS is in its early adolescence, capable of interpreting relatively obvious features in a good EDM, but not yet worthy of attracting serious attention in the protein-crystallographic community. We are currently working toward a more extensive knowledge-based system capable of complete interpretation of medium-quality, medium-resolution (2 to 2.5 Ang.) EDMs. During the next year or two we expect to bring the system's level of performance to the point where it will be noteworthy, by helping in a major way to solve a new protein structure. That goal will be reached, we believe, by incorporating more detailed knowledge about protein chemistry and stereochemistry, and by improving the heuristics for matching the abstracted EDM with structural features determined by the amino acid sequence. We are also extending our preprocessing programs so that more meaningful and/or less ambiguous features can be extracted from the data. Another current activity is a re-design of the system to improve its understandability, modularity and flexibility.

ACKNOWLEDGMENTS

We wish to acknowledge the assistance of our colleagues in the crystallographic community, specifically Dr. Stephan Freer, Dr. Richard Alden and Prof. Joseph Kraut of UCSD, and Dr. Carroll Johnson of ORNL. At Stanford, Ms. H. Penny Nil was instrumental in the initial construction of CRYSALIS; Mr. Eric Grosse assisted in developing algorithms for abstracting the data.

Appendix A

An illustrative example

Space limitations prevent a full example from being included in these proceedings. An annotated typescript, illustrating the performance of CRYSALIS on a typical interpretation problem, will be available at the conference or from the authors upon written request.

References

[Agarwal77]
Agarwal, R. C., and Isaacs, N. W., Method for obtaining a high resolution protein map starting from a low resolution map, Proc. Natl. Acad. Sci. USA, Vol. 74, pp 2835-2839 (1977).

[Engelmore77]
Engelmore, R.S. and Nii, H.P., A Knowledge-Based System for the Interpretation of Protein X-Ray Crystallographic Data, Heuristic Programming Project Memo HPP-77-2, January, 1977. (Alternate identification: STAN-CS-77-589)

[Erman76]
Erman, L. D., Overview of the Hearsay Speech Understanding Research, in *Working Papers in Speech Recognition -IV- The HEARSAY II System*, Carnegie-Mellon University, Computer Science Speech Group, 1976.

[Greer74]
Greer, J., Three-dimensional Pattern Recognition: An Approach to Automated Interpretation of Electron Density Maps of Proteins, *J. Mol. Biol., 82*, 279 (1974).

[Hayes-Roth77]
Hayes-Roth, F. and Lesser, V.R., Focus of Attention in the Hearsay-II Speech Understanding System, *Proc. of 5th Int. Joint Conf. on Artificial Intelligence*, Cambridge, Mass., 1977

[Hermans74]
Hermans, J. Jr. and McQueen, J.E. Jr., *Acta Cryst. A30*, 730 (1974).

[Nii79]
Nii, H.P. and Aiello, N., AGE (Attempt to Generalize): A Knowledge-based Program for Building Knowledge-Based Programs, Proceedings of 6th International Joint Conference on Artificial Intelligence, Tokyo, August 1979.

[Tsernoglou77]
Tsernoglou, D., Petsko, G.A., McQueen, J.E. Jr. and Hermans, J., *Science 197*, 1378 (1977).