Stanford Heuristic Programming Project
Memo HPP-77-2

February 1977

Computer Science Department
Report No. STAN-CS-77-589

A KNOWLEDGE-BASED SYSTEM FOR THE INTERPRETATION OF PROTEIN X-RAY CRYSTALLOGRAPHIC DATA

by

Robert S. Engelmore and H. Penny Nii

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

A KNOWLEDGE-BASED SYSTEM FOR THE INTERPRETATION OF PROTEIN X-RAY CRYSTALLOGRAPHIC DATA

STAN-CS-77-589
Heuristic Programming Project Memo 77-2

Robert S. Engelmore and H. Penny Nii

ABSTRACT

The broad goal of this project is to develop intelligent computational systems to infer the three-dimensional structures of proteins from x-ray crystallographic data. The computational systems under development use 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. The system consists of a number of independent but cooperating knowledge sources which propose, augment and verify a solution to the problem as it is incrementally generated.

KEY WORDS

KNOWLEDGE-BASED SYSTEMS, CRYSTALLOGRAPHY, PROTEIN STRUCTURE, PRODUCTION RULES, REPRESENTATION OF KNOWLEDGE, RULE-BASED CONTROL STRUCTURE.

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2494, Contract No. DAHC 15-73-C-0435, and by The National Science Foundation under Contract No. NSF DCR 74-23461

Table of Contents

TABLE OF CONTENTS

A KNOWLEDGE-BASED SYSTEM FOR THE INTERPRETATION OF PROTEIN X-RAY CRYSTALLOGRAPHIC DATA

ABSTRACT

The broad goal of this project is to develop intelligent computational systems to infer the three-dimensional structures of proteins from x-ray crystallographic data. The computational systems under development use 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. The system consists of a number of independent but cooperating knowledge sources which propose, augment and verify a solution to the problem as it is incrementally generated.

1 Introduction

In this report we present our first investigations into applying Artificial Intelligence methodology to a new task domain, Protein Crystallography. Our goal is to develop an intelligent computational system for inferring the three dimensional structures of protein molecules from x-ray crystallographic and other physical data. Although the computer has for many years been an essential tool in x-ray crystallography research, nearly all its applications have been in the areas of data collection, data reduction, Fourier analysis, graphics and other essentially numerical tasks (Feigenbaum, 1976). Those aspects of molecular structure inference which require symbolic reasoning, and/or which use a significant amount of judgmental knowledge are traditionally performed manually. The structure inference process is basically an iterative cycle of hypothesize, test and refine, of which the first phase (hypothesis generation) involves a significant component of non-numerical analysis.

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. Due to the resolution imposed by the experimental conditions, the electron density distribution is an indistinct image of the structure, which does not reveal the positions of individual atoms. The crystallographer must interpret this function in light of auxiliary data and general principles of protein chemistry in order to derive a complete description of the molecular structure. The ensuing report is devoted to a description of that process, our initial attempts to characterize the process in terms of a knowledge-based problem solving system, and a discussion of the computational system currently being implemented.

2 Description of the problem

The interpretation of an electron density map, derived from the reduction of X-ray crystallographic data, is a necessary and important step in the derivation of the 3-D structure of proteins and other macromolecules. When crystallographers use the term "electron density map" they usually have in mind some pictorial representation of the electron density defined over a certain region of 3-space (usually some fraction of the unit cell of the crystal). The most commonly used representation is a 3-D contour map, constructed by stacking layers of conventional 2-D contour maps drawn on transparent sheets. BY carefully studying the map the experienced protein crystallographer can find features which allow him to infer approximate atomic locations, molecular boundaries, groups of atoms, the backbone of the polymer,

etc. After sever31 weeks (or months) he has built a model of the molecular structure which conforms to the electron density map and is also consistent with his knowledge of protein chemistry, stereochenical constraints 3nd other available chemical and physical data (e.g., the amino acid sequence). A more detailed description of this problem-solving process is given below.

Traditionally, the protein crystallographer embodies his interpretation of the electron density map in a "ball and stick" molecular model, fashioned from brass parts. His task is facilitated by an ingenious device, called a 'Richards box', which permits the model builder to view several layers of the nap through 3 partially transparent mirror, so that the mirror image of his model appears to be "inside" the map. After the model has been completed to the builder's satisfaction, the coordinates of the atoms in the model are recorded, and a process of quantitative refinement begins.

Although many protein structures have been solved in this way, the deficiencies of the brass-model/Richards-box techniques for density map interpretation are well known to those who have used it. Among other difficulties, the 3-D contour map is an awkward representation. The locations of atomic sites and interatomic bonds are seldon directly evident from the contours, at the resolution levels normally obtained. Building a model 'into the density map' is a tedious process of fitting brass parts to regions enclosed by one or more contour levels, a search process which is not very well constrained by the map itself. Another problem is that the brass model sags under its own weight. Consequently the measurement of the coordinates is an errorful process. In recent years an attempt to correct some of these deficiencies has led to the creation of electronic Richards boxes, whereby the model builder can view a CRT display of the electron density map from various angles, and superinpose a line representation of the protein molecule. Although this line of attack is an admirable step towards facilitating the model builder's task, it suffers in two major respects. First, the electron density function is still represented by a contour map. Secondly, the decisions which lead to identification of features in the map are still left entirely to the model builder. The task remains an arduous one of visual pattern recognition, hypothesis generation and testing.

A significant inprovement in automated assistance, beyond those tools mentioned above, would involve a computational system that can generate its own structural hypotheses as well as display and verify them. This capability requires 1)a representation of the electron density function more suitable to machine interpretation, 2)a substantial chemical and stereochemical knowledge base, and 3)a wide assortment of model building algorithms and heuristics, in order to achieve acceptable performance.

In order to obviate the inherent difficulties of contour map

interpretation, investigators are actively pursuing alternate representations. The system under development here is purposely eclectic, exploiting a variety of representations appropriate to an equally varied set of inferential procedures. For example, the skeletal representation of Greer and the ridge line representation of Johnson, discussed in the next section, are both included in our system.

The components of the knowledge necessary for model building fall into three general categories: chemical topology, microstructure and macrostructure. The chemical topology knowledge base is essentially all the known chemical data about the specific protein under study, exclusive of the electron density map itself, e.g., the amino acid sequence, properties of cofactors (if present), and identification of disulfide bridges and/or other special chemical bonds. Microstructural knowledge consists of atomic-level facts about proteins, e.g., the geometry of peptide bonds and amino acid side chains and hydrogen bonding properties. Macrostructure refers to stereotype templates for the plausible major components of the molecules, e.g., alpha helix and pleated sheet, and might also contain statistical correlations linking these stereotypes to the amino acid sequence.

Given these "factual" data and a tractable representation of the electron density map, two more ingredients are required for a complete machine interpretation system. The first is a collection of rules and associated procedures for using this knowledge to make inferences from the experimental data. The second is a problem solving strategy for applying the knowledge sources (KSs) in an effective way, so that the appropriate procedures are executed at the times they will be most productive. Protein crystallographers who build models move continually across a large field of basic facts, special features of the data and implications of the partial model already built, looking for any and all opportunities to add another piece to their structure. There are several requirements to working in this "opportunistic" mode of hypothesis formation: (1) the inference making rules and the strategies for their deployment must be separated from one another, (2) the rules must be separated from the mechanics of the program in which they are embedded, and (3) the representation of the hypothesis space must be compatible with the various kinds of hypothesis generating rules available. (The hypothesis structure represents an a priori established plan for problem solving.) The modularity of such a system allow users to add or change rules for manipulating the data base, as well as to investigate different solution strategies, without having to make major modifications to the system. These issues are discussed further in Sections 6 and 7.

3 Related work

3.1 Protein crystallography

Research on the interpretation of electron density maps has focused on the representation of the electron density function. Greer (1974, 1976) has developed a system for reducing the map to clusters of connected line segments, a process he calls skeletonization. Using the skeletonized map he has developed a set of rules for isolating the main chain, determining directionality and proposing coordinates for specific atoms along the main chain. Greer's program draws heavily on the notion of continuity in the electron density function to produce the skeletonized map, and it uses some knowledge of bond lengths and connectivity to infer main-chain and side-chain coordinates. Knowledge of the amino acid sequence is not assumed. If the sequence were known, the inferences to be drawn from it would presumably be introduced into the program's data base in an ad hoc fashion.

Greer's skeletonization technique, although attractive in its simplicity, suffers in several respects. For one, the procedure is non-deterministic, i.e., one produces a different skeleton by scanning the map in a different order. For another, features of the map easily identifiable to the protein crystallographer, such as helical or ringed structures, are difficult (if not impossible) to identify after skeletonization. The main problem is that one must necessarily lose some information in the process of abstracting a body of numerical data into a highly symbolic representation. One must seek a symbolic representation, or a set of representations, which minimizes the loss of rich detail present in the original data. Skeletonization falls somewhat short in preserving the detail required for complete structure inference.

Recently, another approach toward re-representation of the nap has been to apply numerical analysis to the electron density function. Johnson and Grosse (Johnson, 1976) have developed a method of "ridge-line analysis", wherein they can locate alternating peaks and passes in the electron density function by using an interpolation polynomial. This scheme, which is currently in the implementation and testing phase, will generate a topological representation of the density map, showing all resolved, unique maxima and the most probable interpeak bonds. Although the computational effort required for the application of the interpolation polynomial method is expected to be large, the procedure needs to be done only once for a given structure analysis, and provides both a high level of abstraction of the map and the preservation of most of the significant details that are resolved in the raw electron density function.

3.2 Knowledge-based systems

An area of AI research which the current work resembles is the speech understanding system, Hearsay-II (Erman, 1975)) specifically with respect to the issues of knowledge integration and focus of attention (Hayes-Roth 1976). In Hearsay-II the central task is to build a sentence hypothesis which is a best explanation of the given speech input data. An "iterative guess-building" process takes place, in which a number of different knowledge sources (facts, algorithms, heuristics), operating on various descriptions of the hypothesis, must cooperate. 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 fixed, but depends at any point on what has thus far been established and what available knowledge source is most likely to make further progress. For example, one is unlikely to make much progress by trying to analyze the first segment of the speech wave completely before examining other portions of the utterance. 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 (an illustration of this process in the context of electron density map interpretation is given below). Figure 1 shows the different information levels at which hypotheses are constructed in the **Hearsay-**II system, and some of the knowledge sources used. Knowledge sources are used to establish support for hypothesis elements. These supports are represented by links. A KS may either create, modify or verify a hypothesis element(s) at the target level, given a subset of the existing hypothesis elements at the source level(s). For example, the Syntactic-Semantic Hypothesizer shown in the figure uses syntactic and semantic knowledge of the input language to propose new words adjacent to a word or phrase already on the blackboard.

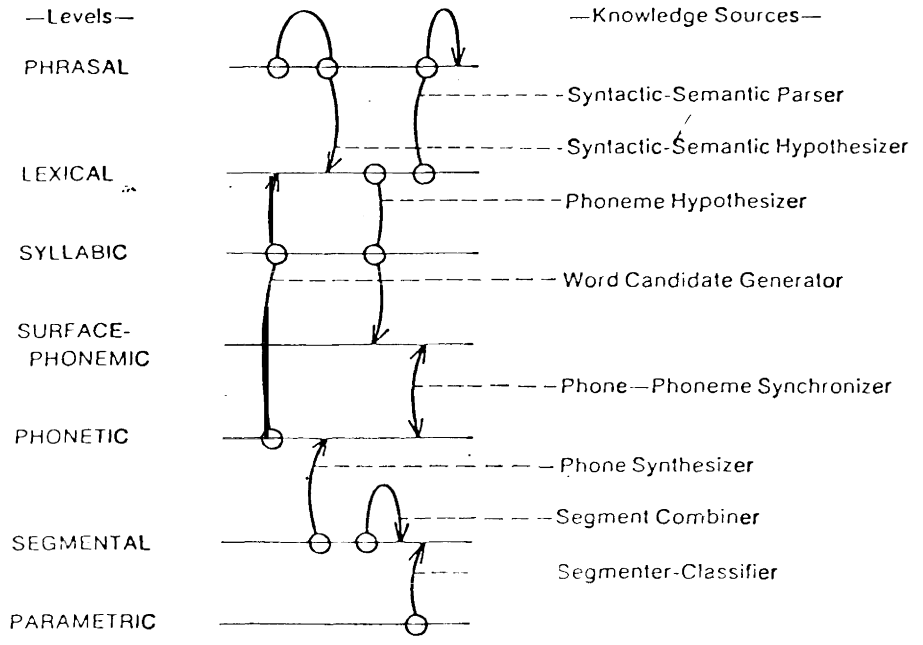


Figure 1. The Current Knowledge Sources in Hearsay II. (from Erman, 1976)

Figures 2 through 4, which are explained in more detail in the next two sections, are descriptions of the protein density map interpretation system. As in Hearsay-II the hypotheses are represented in a hierarchically organized data structure. In our case the different information levels can be partitioned into three distinctly different "planes", but the concept of a globally accessible space of hypotheses is essentially the same for both systems. Knowledge sources also play a similar role as in Hearsay-II, adding, changing, or testing hypothesis elements on the blackboard.

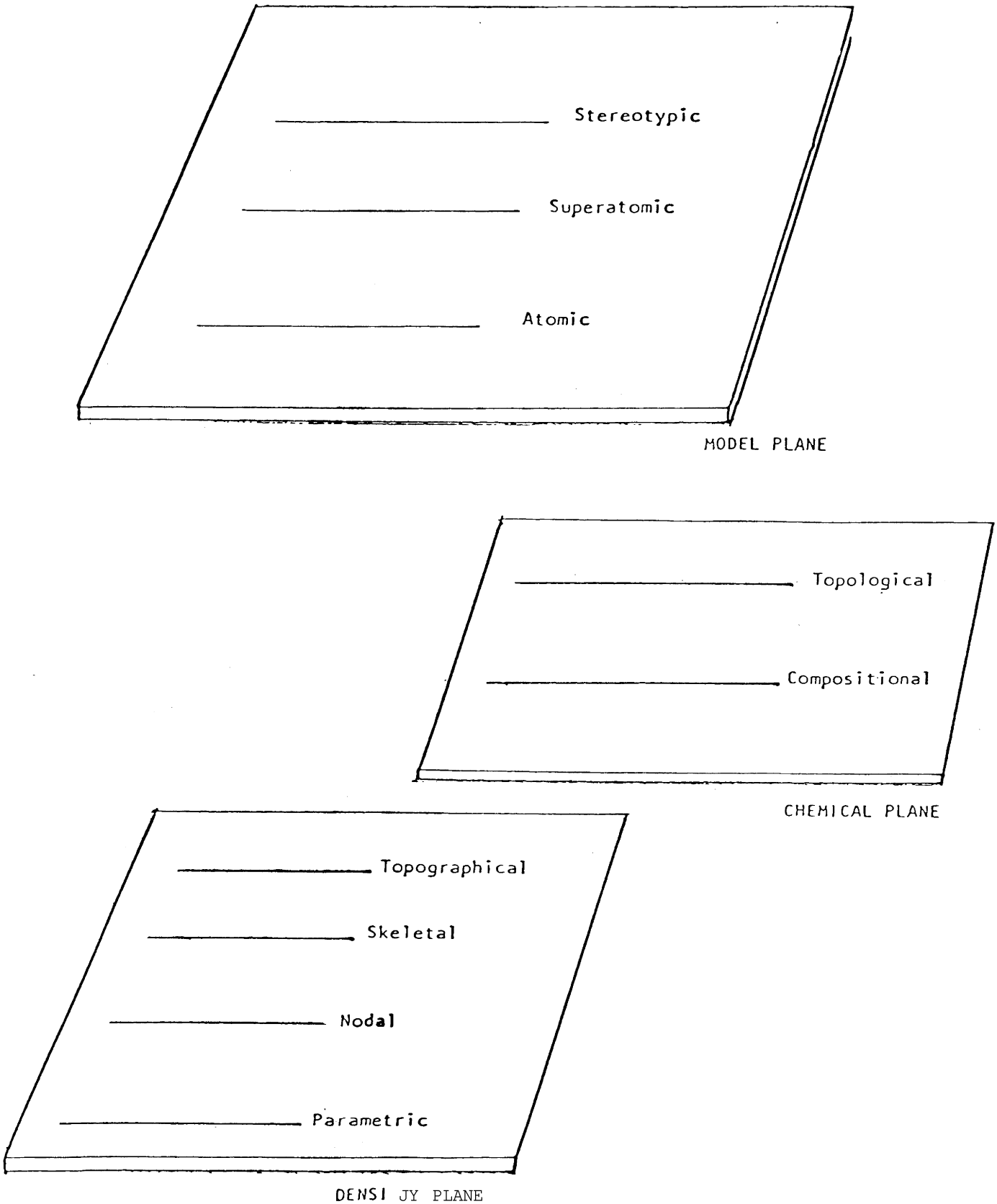


Figure 2. The Space of Hypotheses

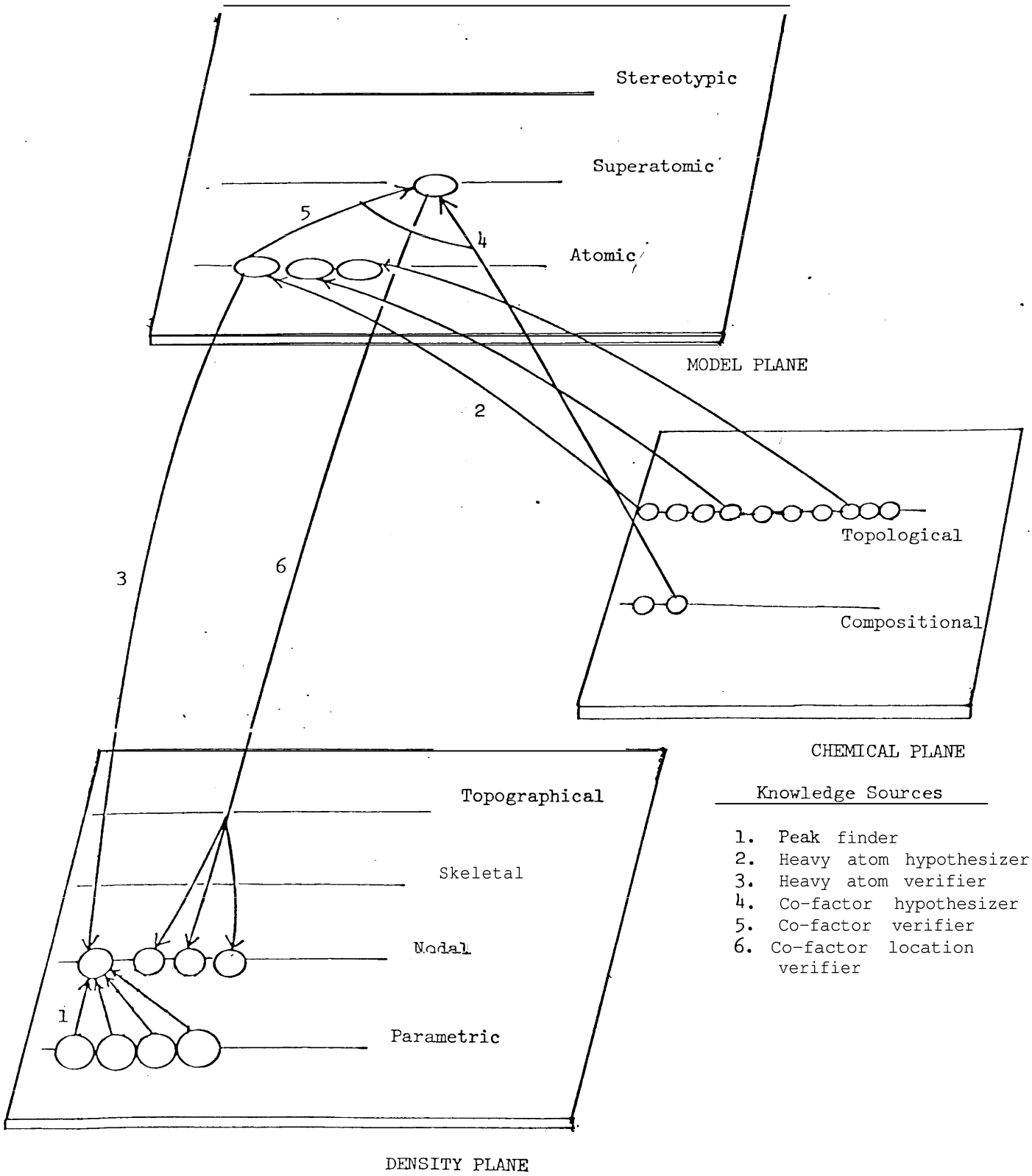


Figure 3. The Application of Knowledge Sources (Ex. 1: Co-Factor Identification)

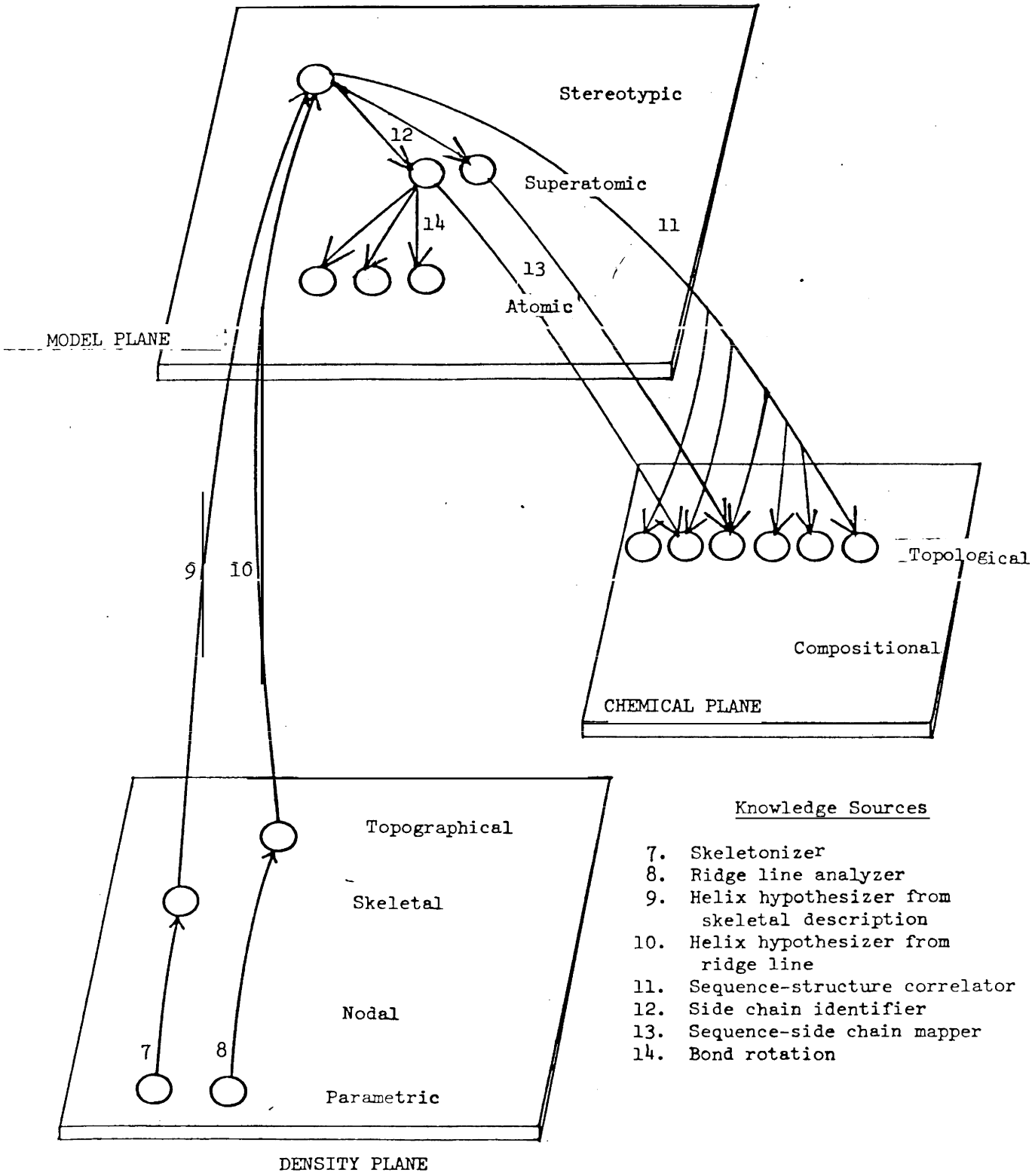


Figure 4. The Application of Knowledge Sources (Ex. 2: Helix Identification)

4 The Nature of a Hypothesis

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 was mentioned earlier, ther are many diverse sources of knowledge being brought to bear on the problem of electron density map interpretation. In order to capitalize on these sources of knowledge, the hypothesis is represented as hierarchically organized levels of descriptions, as shown in Figure 2. A KS is a collection of rules which makes inferences between any two levels in the hypothesis space. There are three levels of description on the model plane. The most detailed level of description of the model is the atomic level; a specification of the spatial coordinates of all atoms in the model with respect to some arbitrary origin (the coordinate of hydrogen atoms are generally omitted). Proteins all 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 the position of the polymeric units (links of the polymeric chain and side chains) 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 in proteins, although they may be associated with a specific family such as the heme binding region in the cytochrome c proteins. This level of description is labelled stereotypic in Figure 2.

A partial or complete hypothesis consists of linked hypothesis elements. A hypothesis element is a labelled node in the space of hypotheses. Attached to each node is a set of attributes which define the hypothesis element in terms appropriate to the level of description on which it resides. For example, each node at the atomic level of description in the model plane corresponds to a discrete atom in the hypothesized structural model. A list of attributes associated with a node of this type includes:

name
type
spatial location (coordinates of the atom)
member of superatom (link to superatoa hypothesis element)
associated peak (link to a density plane description)
associated skeleton node (link to a density plane description)
hydrogen bonds (list of other atoms to which this one is hydrogen bonded)

The Nature of a Hypothesis

Nodes at the superatomic level of description would have a different list of attributes. The relationships between the hypothesis elements are represented by links. For example, a hypothesis element representing a sulfur atom belonging to a particular Cysteine side chain will have a description (ISAMEMBER **CYSi)** attached to it. Another example of a link spanning two levels is (HASASMEMBER **GLUi ALAj** . ..). This could be a description attached to a helix on the stereotypic level indicating a part of the amino acid sequence associated with the helix. There are also relational links confined to a level, such as ISNEXTTO, used to describe the adjacency of the superatoms in terms of the sequence. These links are determined by the **KSs** and represent some of the inferences which they make. The links also have arrowheads to indicate the direction-in which the inferences are being made. For example, if a Cysteine side chain is inferred from a sulfur atom, the link will be from the direction of the atomic level to the superatomic level. On the other hand, if the atomic coordinates of some atoms are inferred from some particular side chain, the links will be from the superatomic to the atomic level. Knowledge sources may make inferences from any level to any other level in Figure 2.

So far we have mentioned the hypothesis structure only with respect to the descriptions of the model. On the other two planes shown in Figure 2 are other descriptions, not of the model but of the data from which the model is derived. The chemical plane contains a static description of known compositional and topological features of the molecule under study; the empirical formula, the amino acid sequence, known hydrogen bonds, di-sulfide bridges, salt links, metal coordinating bonds, etc. These data are errorful and may be modified at a later stage in the structure building process (e.g., an amino acid residue postulated in the sequence may be wrong in light of structural constraints.) However, this occurs rarely and we have, for the time being, made the assumption that the sequence is always correct. Once the amino acid sequence information is assumed to be correct, it can be used as a powerful guide to finding the side chains in the density plane. The use of such knowledge is very similar to the way in which the Syntactic-Semantic Hypothesizer in Hearsay-II uses syntactic and semantic knowledge to predict the next word from the word or phrase already on the blackboard.

The density plane contains the data to be interpreted. In its most elementary form, the density map is typically a very large table of values of the electron density, defined on a **3-dimensional** grid. The number of entries in the table is on the order of 10^5 to 10^6. It is not only prohibitive computationally to search through this data base continually to infer or validate elements of the model. It is also unnecessary, because 1) a large fraction of the map represents regions **outside** the molecules, and 2) we are searching for the positions of 10^3 to 10^4 atoms, so only a fraction of the total table of values contains the most relevant data. Consequently it is clearly

desirable to transform the map to other levels of description which drastically reduce the volume of stored data, yet preserve most of the information required for structure elucidation. Consequently, several other descriptions, or abstractions, of the density map are used. The simplest is a list of peak heights and their locations. Another description exploits the property that most of the protein can be modeled by a single, branched chain, and uses the skeletonization algorithm (Greer, 1974, 1976) to reduce the map to sets of connected line segments. Yet another description is the "ridge-line" representation of the density map, a node-link graph in which the nodes are best estimates of the positions of the maxima, and the links are best estimates of the paths between the maxima (Johnson, 1976).

5 How the Hypotheses are Built by the Knowledge Sources

5.1 Steps in the structure determination process

The inferences made to create, modify or support hypothesis elements are generated by exploiting a large body of facts, formal procedures (algorithms), and informal rules of good guessing (heuristics). These inference makers are called knowledge sources. To appreciate their scope it is instructive to review the steps normally taken by a protein crystallographer in proceeding from an electron density map to a molecular structure. The program organization and the organization of the knowledge sources we have adopted reflect the problem solving processes of the human protein model builder.

There are five major steps in density map interpretation:

A. Qualitative identification
B. Quantitative molecular modeling
c. Calculation of structure factors and comparison with observed structure amplitudes
D. Calculation of a new density map using observed structure amplitudes and model-generated phases.
E. Refinement of the model

Steps C through E, which start with an atomic-level description of the structure, are well-established procedures in crystallographic computing and form the "back end" of a total structure determination system. Our goal is to build the front end, which consists of the first two steps. Qualitative identification is the process of matching parts of the chemical description of the protein (side chain, cofactors, etc.) to corresponding regions of the density map.

Quantitative molecular modeling carries this process further by assigning specific coordinates to the hypothesized structural elements, based on stereochemical or other constraints.

Qualitative identification requires the protein crystallographer to use his knowledge of chemistry and crystallography and his skills in visual identification, all at the same time. In order to develop a program which performs this task automatically, we have analyzed the model builder's reasoning steps in some detail. The process may be subdivided into five sub-processes, although these are not necessarily performed sequentially:

1. Identification of the molecular surface boundary
2. Identification of heavy atoms and major cofactors
3. Identification of the polymer backbone
4. Identification of polymer side chains
5. Identification of minor cofactors and ordered solvent

1. Identification of molecular surface boundary. The size, shape and symmetry elements of the unit cell of the crystal are always known to the crystallographer by the time he has a density map to interpret. He doesn't know, however, where the fundamental repeating unit (i.e., the protein molecule or a cluster of molecules such as a diner, tetramer, etc.) is positioned with respect to the "walls" of his density map. He may thus have, say, the left half of one molecule and the right half of another. For visual identification it is desirable that the map be positioned such that at least one complete and contiguous molecule is contained therein. To accomplish this, the crystallographer uses several sources of information; a) low density regions of the map or "channels" can often be sighted, which indicate the gap between one structural unit and another; b) the molecular weight and volume are used to verify that the hypothesized unit is reasonable in size; c) size and shape data fron light scattering or other auxiliary data may also be used to identify the bounding surface; d) knowledge of the relative densities of the protein and solvent indicate the contrast one may expect between the protein-containing and interstitial regions.

2. Identification of heavy atoms and major cofactor positions (if any are present). The locations of heavy atoms, such as iron, will be obvious in the density map, and are usually the first pieces of structural information to be inferred. Major cofactors often have characteristic shapes, and/or contain the heavy atoms just identified, so they are normally found next. The crystallographer uses the following knowledge sources to carry out this step: a) heavy atoms are located at the maxima in the density map; b) the empirical formula of the protein tells him how many and of what type of heavy atoms and cofactors to look for; c) the number of disulfide bridges, determined from chemical analysis, is used to direct the search for these peaks in

the density map; d) the atomic numbers of the atoms determine relative peak heights, so that different types of heavy atoms may be distinguished; e) the known shape of major co-factors is used to direct the search for their positions in the map (e.g., a flat, quasi-circular group).

3. Identification of the polymer backbone. Distinguishing the main chain of the protein from side chains and cofactors is a crucial task in the model building process. The relevant knowledge sources here include: a) if a relatively long connected region in the density map can be identified, it usually indicates the image of the main chain; b) the number of amino acids in the protein implies a total length for the main chain; c) the amino acid sequence, including disulfide bridges, can be used to infer the length of loops in the chain, d) predictions of the fraction of the polymer which is in a helical configuration can be obtained fron optical rotatory dispersion data or from statistical analyses of amino acid sequences in known proteins (Chou, 1974); e) knowledge of the geometry of characteristic configurations, such as the alpha helix or the pleated sheet, can be used to match their shapes against clusters of density in the map.

4. Identification of polymer side groups. Identifying even one or two specific side chains along the polymer allows the model builder to start matching his model to the amino acid sequence. Once this foothold is established, he can make rapid progress in adding the side chains to the backbone, because he has strong expectations which limit the possibilities. Among the many knowledge sources employed for this task are: a) protrusions found on the backbone at regular intervals indicate the presence of side chains and their points of attachment; b) the "4 Angstrom" rule for alpha carbon separation can be used to verify the points of attachment of the side chains; c) the sizes and shapes of these bumps can be used to infer which amino acid side chains it may represent - e.g., big, flat bumps are most likely to be phenylalanine, tyrosine or arginine; d) the amino acid sequence, particularly useful when two or more adjacent side chains can be identified, e) the shapes of the amino acid side groups can be used to verify an identification of a side chain in the map; f) family resemblances among classes of proteins can be exploited to locate relatively long sequences in the density map; g) special properties of the different amino acid residues are also used, such as their tendencies to occur within or outside of helical regions, or their tendencies to point away from (hydrophobic) or toward (hydrophilic) the surface of the molecule.

5. Identification of minor cofactors and ordered solvents. Small clusters of atoms often co-exist with the protein, and it is necessary to distinguish them as separate entities. Examples are the inhibitor in an enzyme-inhibitor complex, or interstitial water molecules. Information sources for this phase of the analysis include:

a) the residual density in the map; b) the empirical formulae for the cofactors and the solvent molecules; the general rules that c) the solvent is almost always located outside the molecular boundary; d) substrate/inhibitor cofactors have access to both the inside and the outside; e) the ordered solvent is usually hydrogen-bonded to polar side chains.

5.2 How the automated interpretation system uses knowledge - Examples

We have begun building a system which employs those knowledge sources used by the crystallographers which are relatively easy to implement. The system's control structure (see Section 7) permits the knowledge sources to be discrete, independent entities, so that the addition of new knowledge sources, or new rules within the KSs, involves little or no reprogramming of the existing system. Which knowledge sources are used, and in what order, is determined by the latest changes in the hypothesis. In addition, the complete hypothesis space is always available for pursuing other strategies.

Two examples are given here which illustrate the use of several knowledge sources and their integrated effects. The first is a subproblem which the current system can solve, and, though relatively trivial, demonstrates the flavor of the system's problem-solving behavior. The second is a more difficult subproblem but also a more typical model-building task.

5.2.1 Example 1 (see Figure 3)

The knowledge sources used in the first example are shown schematically in Figure 3. The problem is that of cofactor identification, step 2 in the above discussion of qualitative identification. In this example the structure under investigation was a member of the cytochrome c family of proteins. The density map was derived from a theoretical model of the protein, not from crystallographic data, so the density map is of high quality. The electron density function was computed to a resolution of 2 Angstroms and sampled on a grid of approximately 1 Angstrom spacing. Consequently most atoms in the structure are not individually resolved in the map. The most readily identifiable features in the map are the heavy atoms -- iron and sulfur -- and the heme group, characteristic of all members of this protein family.

The program starts with the density map, the composition of the protein, the amino acid sequence, and the general knowledge base discussed previously. As shown in the figure, six knowledge sources

are invoked. KS-1 is a preprocessor which abstracts from the parametric description of the density map (i.e., the lattice-sampled electron density function) a list of the locations of the most prominent peaks, sorted from highest to lowest peak heights. Thus several points in the parametric representation, in the vicinity of a peak, are mapped into a single hypothesis element at the nodal level, as shown. Each element at the nodal level is assigned a name, and its height and position are entered as properties of that nane. KS-2 infers from the chemical data that certain heavy atoms are present in the structure. For example, the cysteine side chains at positions 14, 17, 55 and 91 in the sequence are noted and, using the global knowledge base, infers that there are four heavy atoms of type sulfur in the protein. A similar inference can be made for the one iron in the protein. KS-2, therefore, creates and establishes support for several heavy atom hypothesis elements at the atomic level of the model description. These elements are assigned identifiers (A1, A2, etc.) and properties which associate them with specific atoms in the topological description are attached. KS-3 establishes the spatial locations of tlie atoms by looking at the list of nodes and selecting candidates which are most likely to correspond to the heavy atoms. The iron atom position is taken as the position of the highest peal: in the map . The sulfur atoms in the vicinity of the iron are also located in the node list, using general knowledge of the cytochrome c family structure.

Having inferred as much as possible about heavy atoms at this stage of the analysis, the system shifts its attention to locating the heme structure. KS-4 makes the simple inference, based on the protein's family membership, that one of the superatomic hypothesis elements is a heme, and creates that element on the "blackboard". KS-5 provides support for the heme by linking it with the iron atom already found. The combination of having located the iron atom and having hypothesized the heme superatom triggers the heme locater, KS-6. KS-6 searches through the node list to find those peaks in the density which are most likely to lie within the planar structure of the heme, and predicts the direction of the normal to the plane. We present here a trace of the first few steps of the program's reasoning activity for this example in order to illustrate the flow of control as it evolved. The terminal output is given immediately below. Annotations occur within the output in lower case type, and also occur following the output.

```
INITIAL VALUES FOR CYTOCHROME_C2
COFACTOR:  HEME
KNOWN.LOCATIONS:  ((FE .216 .063 .427))
SEQUENCE:  GIVEN

  IMFERENCE:  EVENT-1  BY RULE-1  IN RULESET  INITIALIZATIONRULES

     EVENT NAME: COFACTOR-POSITED
     CURRENT HYPOTHESIS ELEMENT: SA1
     NEW PROPERTIES:  ((TYPE COFACTOR) (NAME HEME))
```

A set of rules, called "initializationrules", is called unconditionally in order to "get something on the board". Here the first hypothesis element is created in the model plane, and the token "cofactor_posited" becomes the initial item on the event list.

```
  INFERENCE:  EVENT-2  BY RULE-1  IN RULESET  INITIALIZATIONRULES

     EVENT NAME: HEAVYATOM POSITED
     CURRENT HYPOTHESIS ELEMENT: A1
     NEW PROPERTIES:  ((TYPE FE) (NAME FE) (BELONGSTO HEME)
                       (MEMBEROF SA1))
```

The same rule may generate more than one event. Here the rule which just posited a heme structure in the protein also creates a subsidiary hypothesis (the iron atom) and establishes membership links between the two hypothesis elements. (This inference was made using general knowledge about the composition of the heme group.) Associated with each event is a particular hypothesis element, which is the current focus of attention. The event may signal the creation of the hypothesis element, as it does here, or may signal the establishment of new properties for a pre-existing hypothesis element, as in the next event below.

```
  INFERENCE:  EVENT-3  BY RULE-2  IN RULESET  INITIALIZATIONRULES

     EVENT NAME: HEAVYATOM LOCATED
     CURRENT HYPOTHESIS ELEMENT: A1
     NEW PROPERTIES:  ((SPACE-LOC (.216 .063 .427)) (D-NODES (ND1)))

  INFERENCE:  EVENT-4  BY RULE-4  IN RULESET  INITIALIZATIONRULES

     EVENT NAME: HEAVYATOM POSITED
     CURRENT HYPOTHESIS ELEMENT: A2
     NEW PROPERTIES:  ((TYPE S) (NAME SG14) (BELONGSTO (CYS 14)))

  INFERENCE:  EVENT-5  BY RULE-4  IN RULESET  INITIALIZATIOMRULES

     EVENT NAME: HEAVYATOM_POSITED
```

```
     CURRENT HYPOTHESIS ELEMENT: A3
     NEW PROPERTIES: ((TYPE S) (NAME SG17) (BELONGSTO (CYS 17)) )

  INFERENCE: EVENT-6    BY RULE-4  IN RULESET  INITIALIZATIONRULES

     EVENT NAME : HEAVYATOM_POSITED
     CURRENT HYPOTHESIS ELEMENT: A4
     NEW PROPERTIES: ((TYPE S) (NAME SD55) (BELONGSTO (MET 55)))

  INFERENCE: EVENT-7    BY RULE-4  IN RULESET  INITIALIZATIONRULES

     EVENT NAME : HEAVYATOM_POSITED
     CURRENT HYPOTHESIS ELEMENT: A5
     NEW PROPERTIES: ((TYPE S) (NAME SD91) (BELONGSTO (MET 91)))
```

Events 4 thru 7 were generated by a rule which scans the amino acid sequence for those side chains that should be "visible" as heavy atoms in the density plane. These heavy atoms would then serve as foci of attention for further hypothesis formation activities.

```
EVENT-1   COFACTOR-POSITED   SA1
```

The normal processing cycle begins here. An event is picked off the event list, here identified by its number, name and associated hypothesis element. In the current implementation the event list is a queue, so that the first event generated is the first to be examined. The event is passed first to the strategy rule processor to see if any special strategies apply. In this case, a strategy rule for merging two events (1 and 3) does apply, and a new event is placed in the front of the event list, overriding the breadth first strategy represented by the queueing of events.

```
MERGED INFERENCE: EVENT-8  FROM  EVENT-1  AMD  EVENT-3
    BY STRATEGY RULE-1

EVENT-8   HEME_AND_FELOC   SA1

  INFERENCE: EVENT-9    BY RULE-1  IN RULESET  HEMEANALYSIS

     EVENT NAME: HEME LOCATED
     CURRENT HYPOTHESIS ELEMENT: SA1
     NEW PROPERTIES: ((D_NODES (ND17 ND30 ND33 ND38)))
```

The new "merged" event is passed down to the event processor, which matches the event name to a rule set called "hemeanalysis". A member of this rule set is found to be applicable, thereby establishing new properties for the current hypothesis element, and a new event is queued on the event list.

```
EVCIJT-2   HEAVYATOM_ POSITED  A1
```

```
INFEKENCE:   EVENT-10     BY RULE-1   IN RULESET   FINDHEAVYATOMS

   EVENT NAME: HEAVYATOM LOCATED
   CURRENT HYPOTHESIS ELEMENT: A3
   NEW PROPERTIES: ((D.NODES (ND3))
                    (SPACE-LOC  (.3425 .0917 .4778)))

INFERENCE:   EVENT-11     BY RULE-1   IN RULESET   FINDHEAVYATOMS

   EVENT NAME: HEAVYATOM  LOCATED
   CURRENT HYPOTHESIS ELEMENT: A2
   NEW PROPERTIES: ((D.NODES (ND2))
                    (SPACE_ LOC  (.1649 -.0868 .4673))
```

Event-2 now comes to the top of the list, and triggers a new ruleset, called "findheavyatons". The application of this knowledge source results in establishing links between the two hypotheses elements, A2 and A3, and specific peaks in the density map.

The event processor is governed by its own set of rules. If an event triggers a set of knowledge rules, and no inferences can be made, the failure is due either to insufficient data, a lack of necessary information in the model thus far constructed, or ignorance of that particular knowledge source. Since the model hypothesis may change as the result of processing other events, the event is placed on the job-list, to be examined at a later time by other knowledge sources. Another type of failure may be due to general ignorance, i.e., the program simply has no knowledge sources which may be invoked for the current event. An event rule for this situation is to place the event at the back of the event-list, awaiting either the creation of new events which may be merged with the current one to form a "processable" event; or the addition of new knowledge sources to the system.

5.2.2 Example 2 (see Figure 4)

The second example is the subproblem of helix identification. The model builder attempts to find helical regions in his density map at an early stage in the model building process, because such regions have a well-defined density in the 3-D contour map. A helix of sufficient length (at least seven residues) will appear in the map as a "rod" of high density, often with a hole running through it. Once the helix template has been fitted into the density, the model builder can exploit its highly constrained structure to determine the direction of the chain, the regions of surrounding density which correspond to side chains attached to the helix, and the identity of those side chains having recognizable sizes or shapes.

The corresponding analysis made by the automated system is sketched in Figure 4. In the density plane, the density map is abstracted into either a skeletal description or a ridge line representation (KS-7 and KS-8, respectively), as discussed previously. KS-9 examines the shape of the main chain hypothesized by the skeletonizer and looks for helical features (e.g., patterns formed by vectors between adjacent carbonyl groups). KS-10 is a similar knowledge source which uses the more detailed representation of the density function provided by the ridge line analysis. If either KS is successful an hypothesis element is entered at the stereotypic level on the model plane. Properties for this element include the location of its centroid, the direction of the helical axis, number, size and shape of side chains, and polarity. KS-11, the sequence-structure correlacor, examines the amino acid sequence and predicts subsequences which are' likely to be within helical regions. KS-12 uses the side chain information associated with the helix to establish hypothesis elements at the superatomic level, one for each side chain. KS-13 matches the side chain sizes and shapes with those expected in the helical subsequences in order to establish the identity of these superatoms. KS-14 creates hypotheses at the atomic level from the known superatoms by determining the appropriate translation and bond rotations which bring the side chain template for the current superatom hypothesis into best agreement with peak locations in the density map.

6 Representation of Knowledge in the System

As illustrated in the previous section there are many diverse sources of information used in protein structure inference. The problem of representing all this knowledge, in a form which will allow it to be used cooperatively and efficiently in the search for plausible hypotheses, is of central concern to this research. The system currently under development draws upon many concepts which have emerged in the design of other large knowledge-based systems, e.g., the use of production rules and blackboards. In this section we describe how these concepts have been adapted to our particular task.

Knowledge consists of facts, algorithms and heuristics (rules of good guessing). Facts required for protein structure inference are general physical, chemical, stereochemical and crystallographic constraints. Typical factual knowledge stored in the system includes physical properties of the elements commonly found in proteins, molecular structure and chemical properties of the twenty amino acids, bond lengths and symmetry properties of various crystal structures. These facts are encoded as tables or in property lists attached to specific structural entities. An example of the latter is the property list associated with glutamic acid, shown in Figure 5. Factual

knowledge comprises a global data base, which is used as needed by the knowledge sources as they attempt to infer elements of the structural hypothesis.

GLU

FULL N AME	GLUTAMIC_ACID
POLARITY	ACIDIC
IIYDRO	HYDROPHILIC
H BOND ACCEPTOR	(6 (OE1 . 3) (OE2 . 3))
H_BOND_DONOR	NIL
SHAPE	ACYCLIC_BRANCHED
RESIDUE-WT	72.0
HELIX	1.53
BETA	0.26
ATOM_LIST	((CA 0.0 0.0 0.0) (CB -.05 -.933 1.244) (CC 1.221 -1.754 1.546) (CD 1.431 -3.015 .625) (OE1 .957 -3.081 -.47) (OE2 2.13 -3.821 1.239))
BOND_LIST	((CB . CG) (CC . CD) (CD . OE1) (CD . OE2))
SEGMENTATION_LIST	(BO (B1 (B2 B3 B4)))

Figure 5. A Component of the Global Data Base: Property List for Glutamic Acid

Algorithms and heuristics comprise the formal and informal knowledge which generate and/or verify hypothesis elements. We have been guided by two general principles in the representation of the knowledge sources:

1) decompose identifiable areas of knowledge into elementary units, each of which increments the hypothesis when specified preconditions are met.

2) represent the elementary units as situation-action rules.

To illustrate, consider the relatively simple example of heavy atom location. This subproblem is decomposed into two independent parts: 1) inferring the presence of heavy atoms and 2) determining their spatial locations. These two independent parts are represented as two separate KSs, invoked under different conditions. In the specific example of cytochrome c2, the presence of the heavy atoms is

inferred from a KS containing two rules, one which infers the iron from the presence of the heme cofactor in the composition list, and the other which infers the presence of sulfur atoms from the amino acid sequence. The two rules nay be stated as situation-action rules as follows:

```
Rule 1
   IF the composition list contains a cofactor of type heme,
   THEN :
     1) create a superatom node of type hene in the model plane,
     2) create an atom node of type iron in the model plane,
     3) create membership links between the iron and the heme,
     4) put "cofactor_posited" on the event-list,
     5) put "heavyatom_posited" on the event-list.

Rule 2
   IF the amino acid sequence is given,
   THEN:
     for each residue in the sequence,
     1) IF the residue is cysteine,
        THEN:
           1.1) create an atom node of type S in the model plane and
                name SGn, where n is the sequence no. of the residue,
           1.2) put "heavyatom_posited" on the event-list;
     2) IF the residue is methionine,
        THEN :
           2.1) create an atom node of type S in the model plane
                and name SDn,
           2.2) put "heavyatom_posited" on the event-list.
```

Note that in both rules above several actions may be performed for a given situation. Also, as shown in rule 2, an action may itself be a situation-action rule, and may be iterative. Not shown here, but present in the LISP implementation of these rules is a position in the rule for setting parameter values, to avoid repetitious calculation of parameters appearing in several situation-action clauses. Also note that at least one of the actions of each rule is to place a token on an event-list. In the actual implementation the syntax of the "action" clause is represented as one function. An example follows:

```
syntax:  (<inference type> (element being changed> <att-value pairs>)

example: (HEAVYATOM.POSITED (GENATOM) ((TYPE FE) (CELONGSTO HEME1))
```

In this example, the hypothesis element A1 will be created. It will be described as an iron atom belonging to a heme. Further, an event HEAVYATOM.POSITED will be generated and queued on the event list. The event-list is used by the interpreter, discussed in the next section,

to determine what to do next, i.e., which set of knowledge sources will be invoked after the current event has been processed.

7 Control Structure for the Map Interpretation System

7.1 Event-driven versus goal-driven control

There are several choices of control structure faced by the designer of a knowledge-based system. Basically the choices are among points on a spectrum, at the extremes of which are goal-driven and event-driven systems. In a goal-driven system (of which MYCIN is a well-known example (Shortliffe, 1976)) the rule interpreter selects a rule which concludes with the goal being sought. In our system, we might imagine having such a goal rule as follows:

IF
1) the topological description is complete, and
2) the coordinates of all atoms in the structure are assigned, and
3) the structure satisfies stereochemical constraints, and
4) the structure is consistent with the electron density function, and
5) the structure is consistent with auxiliary chemical data,

THEN:
signify that a model has been completed.

The interpreter would then attempt to verify each of the premises in the goal rule. To do that, other rules would be selected whose conclusions (the right-hand sides) verified the premises under consideration and the interpreter would attempt to verify the premises of these rules, and so on, working through the list of rules in this recursive fashion. The program's focus of attention is determined by the current rule whose premises are being evaluated. Many levels of recursion may occur before a rule is reached which is relevant to the current state of the system. A goal-driven monitor is attractive, in that it pursues a logical chain of reasoning, in which the purpose of each move is clearly revealed by the tree of subgoals.

An alternate way to focus attention is to employ an event-driven control structure. In this scheme the current state of the hypothesis space determines what to do next. The monitor continually refers to a list of current events - the event-lists mentioned In the rules discussed above - which is used to trigger those knowledge

sources most likely to make further headway. As a knowledge source makes a change in the current hypothesis, it also places a symbol on the event-list to signify the type of change made. Thus as events are drawn from the event-list for processing, new events are added, so that under normal conditions the monitor always has a means for choosing its next move.

The system we are currently developing operates in both goal-driven and event-driven modes, with an emphasis on the latter. 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-lists. Thus 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 as the primary mode of operation is based partly on efficiency in selecting appropriate knowledge sources and partly on conformity with the structure modeling process normally employed by protein crystallographers. Some parts of the model building process, however, are handled more appropriately within a goal-driven framework. For example, having identified a side chain within a particular region of the electron density map, it may be desirable to defer the task of determining the locations of the constituent atoms in that side chain until other, neighboring side chains have also been located. The system then sets up a subgoal (find the atomic positions of superatom SA17) and places it on a list of jobs. Whether to process this subgoal or not is determined by the strategy rules which take into consideration the impact of pursuing this subgoal on the overall solution and the likely success of such a move.

7.2 Knowledge-deployment rules, event rules and strategy rules

The formal and informal procedures which comprise our knowledge sources are expressed as rules, as discussed above. These rules are collected into sets of rules, each set being appropriate to use on a particular class of events. The events generally reflect the level on which the inference is being made, which in turn reflects the level of the detail of the model. The correspondence between event classes and rule sets is established by another set of rules, the event rules. The event rules thus form a second layer of rules which direct the system's choice of knowledge sources for a given event, reflecting the system's knowledge of what it knows. (A similar set of rules, the job rules, perform the same role when the system operates in goal-driven mode.) Maintaining the rule-based structure affords a flexibility in choosing different combinations of knowledge sources to work together, without having to make any changes in the knowledge sources themselves. Thus,

yet a higher level knowledge source, the strategy rules, can manipulate the events in order to choose the appropriate combination of KSs suited to a particular stage or state in the solution hypothesis. This was illustrated in Example 1 when two events were merged into one event by a strategy rule.

The part of the monitor which interprets and obeys the event rules may be likened to a middle-level project manager, who knows which specialists to call in as new, partial solutions to a particular problem are discovered. Continuing the analogy, the middle-level manager occasionally gets stuck and needs help from a higher level of management. As mentioned earler, some high-level decision, such as merging two or more events to produce a new event that can lead to further progress, or shifting from event-driven to goal-driven mode, is required. This level of decision making is embodied in a set of strategy rules, which are used for directing the top level flow of control. We thus have a completely rule-based control structure, employing three distinct levels of rules (or knowledge): the specialist, commonly called the knowledge sources, the event processing rules (or job processing rules), representing knowledge about the capabilities of the specialist, and the strategy rules which know when to use all available knowledge to solve the problem. Although this pyramidal structure of rules and meta-rules could continue indefinitely, the flexibility of knowledge deployment offered by our three-tiered system would appear to be sufficient for this problem solving system. Similar ideas in a simpler context have been explored by Davis (1976) for the MYCIN system.

8 Summary

In this report we have attempted to describe, in all its complexity, the problem of determining the structure of proteins. Conventional methods for solving this problem demonstrate that many kinds of formal and heuristic knowledge cooperate in building the structural hypothesis, piece by piece. A characteristic feature of the process is that a contribution by one KS often enables other KSs to build further. We have also described a knowledge-based system, now under development, which we feel is suited to the activities involved in this opportunistic way of solving problems.

ACKNOWLEDGEMENTS

We gratefully acknowledge the substantial contributions made by

Summary

Carroll K. Johnson, both in the initial formulation of the problem discussed here and in providing model-building expertise. We are also indebted to Prof. Joseph Kraut, Dr. Stephan Freer, Prof. Ng. Xuong and other members of the protein crystallography group at UCSD for their help in formulating strategies for interpretating protein electron density maps.

9 References

Chou, P. Y. and Fasman, G. D. (1974), Biochemistry 13, 211-245.

Davis, R. (1976), "Applications of Meta Level Knowledge to the Construction, Maintenance and Use of Large Knowledge Bases", Stanford Artificial Intelligence Laboratory Memo AIM-283, 1976.

Ernan, L. D. (1976), "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.

Feigenbaun, E. A., Engelmore, R. S. and Johnson, C. K., "A Correlation Between Crystallographic Computing and Artificial Intelligence Research", submitted for publication in Acta Crystallographic.

Greer J. (1974), J. Mol. Biol. 82, 279-301.

Greer, J. (1976), J. Mol. Biol. 100, 427-458.

Hayes-Roth, F. and Lesser, V.R., "Focus of Attention in a Distributed-Logic Speech Understanding System". Proc. of IEEE, Int. Conf. on ASSP, Philadelphia, Pa., 1976.

Johnson, C. K. and Grosse, E. (1976), Abstract, Proceedings of the American Crystallographic Association, Evanston, Ill., August, 1976.

Shortliffe, E. H. (1976), Computer-Based Medical Consultations: MYCIN, Artificial Intelligence Series 2, Elsevier, New York, 1976.

10 Appendix. A Glossary of Terms Used in Protein Crystallography

PROTEIN

A linear chain of amino acids. Of the several classes of proteins, the most interesting are the enzymes, which have a generally globular shape . Proteins are often described as a polypeptide chain plus amino acid residues, or side chains, attached at each link in the chain.

POLYPEPTIDE

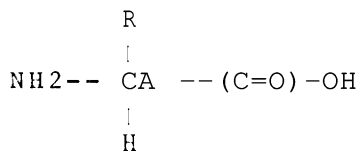
A repeating sequence of atoms,

$$--CA--(C{=}O)--NH--CA--(C{=}O)--NH--CA--(C{=}O)--NH--$$

where CA is the alpha carbon to which the amino acid residue is attached.

AMINO ACID, AMINO ACID RESIDUE

An amino acid has the following topological structure:



The alpha carbon (CA) is surrounded by an amino group, a carboxylic acid group, a hydrogen atom, and a side chain (R) which characterizes the particular amino acid. By removing a molecule of water (H on one side, OH on the other) the remaining amino acid residue can be linked to other amino acid residues in a polypeptide chain (q.v.). There are twenty common amino acid residues found in proteins. They are referred to by either their full names, their 3-letter names, or their 1-letter names, as follows:

1.	ALANINE	ALA	A
2.	ARGININE	ARN	R
3.	ASPARAGINE	AS N	N
4.	ASPARTIC ACID	ASP	D
5.	CYSTEINE	CYS	C
6.	GLUTAMIC ACID	GLU	E
7.	GLUTAMINE	GLN	Q
8.	GLYCINE	GLY	G
9.	HISTIDINE	HIS	H
10.	ISOLEUCIME	ILE	I
11.	LEUCINE	LEU	L
12.	LYSINE	LYS	K

13.	METHIONINE	MET	M
14.	PHENYLALANINE	PHC	F
15.	PROLINE	PRO	P
16.	SERINE	SER	S
17.	THREONINE	THR	T
18.	TRYPTOPHAN	TRP	W
19.	TYROSINE	TYR	Y
20.	VALINE	VAL	V

The PRIMARY STRUCTURE of a protein is a description of the amino acid sequence.

The SECONDARY STRUCTURE of a protein is a description of the structure in terms of common substructures, such as alpha helices and pleated (or beta) sheets.

The TERTIARY STRUCTURE is a complete specification of the positions of all atoms in the molecule.

ALPHA HELIX

A special configuration of the polypeptide chain, similar to the helical construction of DNA and RNA. There are approximately 3.6 alpha carbons per complete turn of the helix. The helix is held in place by hydrogen bonds between the backbone nitrogen and the carbonyl oxygen four links further down the chain. The protein myoglobin has a high helix content.

PLEATED SHEET or BETA SHEET

The polypeptide chain can often make a U-turn and run back alongside itself, locking the two chains together by hydrogen bonding. Pleated sheets can be either parallel or anti-parallel. Silk is an example of a protein which is almost entirely in the pleated sheet configuration. The globular protein concanavalin A (a toxic protein from jack beans) has a high beta sheet content.

CO-FACTOR

A co-factor is an integral part of the protein, although it is not part of the sequence of amino acids. The heme group in the globin and cytochrome families is an example of a co-factor. Co-factors are held in place by hydrogen bonds or metal coordination bonds to the amino acids in the polymeric sequence.

HYDROGEIJ BOND

A hydrogen link between two other atoms,

i.e., $X^{--}H...Y$ where X,Y = O,N

Appendix. A Glossary of Terms Used in Protein Crystallography

COORDINATION BOND

A bond of the sort metal--X where X = O,N and metal = Fe,Cu,etc.

DI-SULFIDE BOIJD

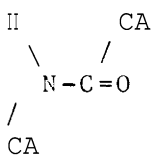
I.e., -S--S-

VAN DER WAAL'S RADIUS

The effective radius of an atom, determining the distance of closest approach of two non-bonded atoms.

AMIDE PLANE

Between every pair of alpha carbons in the polypeptide chain are two groups, -NH- and -(C=O)-. The atoms of these two groups, plus the two alpha carbons, all lie in a plane, called the amide plane.



Amide Plane

DIHEDRAL ANGLES

Angles between planes containing atoms. A pair of dihedral angles which specify rotations about the CA--N and C--CA bonds determines the orientation of one amide plane with respect to an adjoining amide plane. The configuration of the protein backbone is thus completely specified by a list of dihedral angle pairs, one pair for each set of adjacent amide planes, assuming a fixed geometry for the amide planes.

UNIT CELL

The basic repeating parallelepiped in a crystalline structure. The crystal can be "generated" by translating the unit cell along each of its three principal axes.

SYMMETRY ELEMENT

A geometrical entity, such as a point, a line, or a plane, with respect to which a particular symmetry operation is performed.

Appendix. A Glossary of Terms Used in Protein Crystallography

SYMMETRY OPERATION

The actual or hypothetical movement of a body, by translation, rotation (an n-fold rotation is a rotation of 360/n degrees, where n=2,3,4,or 6), rotatory inversion (rotation plus inversion of all points through a center lying on the axis of rotation), screw rotation (rotation plus translation along axis by 1/n of unit cell dimension) or translation plus reflection (glide plane operation). Successive applications of a symmetry operation must eventually return the object to its initial position (or, in a crystal, to one related by translation). Since proteins are inherently left-handed, symmetry operations involving reflections or inversion are prohibited.

POINT GROUP

A group of symmetry operations, all of which leave unmoved one point within the object to which they apply. The kinds of symmetry elements that may be present include simple rotation and rotatory-inversion axes; the latter include the center of symmetry and the mirror plane. Since one point remains invariant, all rotation axes must go through this point and all mirror planes must contain it. A point group is used to describe isolated objects such as single molecules.

SPACE GROUP

A group or array of operations consistent with an infinitely extended, regularly repeating pattern. There are just 230 three-dimensional space groups, which can be obtained by the addition of translation components to the 32 point groups appropriate for structures arranged on lattices. The additional symmetry elements present in space groups include simple translations, screw axes, and glide planes.

TRIAL STRUCTURE

A possible structure for a crystal, which is tested by a comparison of calculated and observed structure factors and by the results of an attempted refinement of the structure.

FOURIER DENSITY MAP

The electron density function for a crystal sampled at a set of three-dimensional grid points. This map is calculated as a three-dimensional Fourier series using the structure factors as coefficients.

Appendix. A Glossary of Terms Used in Protein Crystallography

STRUCTURE FACTOR (F)

The magnitude of the structure factor, |F|, is the ratio of the amplitude of the radiation scattered in a particular direction by the contents of one unit cell to that scattered by a single electron under the same conditions. The structure factor has both a magnitude (amplitude) and a phase; from the intensity we can derive directly the amplitude but not the phase. Structure factors represent values, at the reciprocal lattice points h, k, 1, of the Fourier transform of the electron distribution in one unit cell. The structure factor depends on:

1. the nature of the scattering material
2. the arrangement of the scattering material (including thermal motion)
3. the direction of scattering.

The experimentally measured ("observed") structure factor amplitudes are designated by |Fo|; those calculated for a model of the structure are designated |Fc|.

INTENSITY (I)

The calculated or experimentally measured quantity related to the structure factor F:

$$I = |F|^2 * \text{geometrical correction factor}$$

AMPLITUDE

The modulus of the structure factor, i.e. |F|.

PHASE

The quantity phi in the identity F = |F|*exp(phi)

THE PHASE PROBLEM

Given all the experimentally measured values of |F|, find the F's so that the Fourier density nap can be calculated.

ISOMORPHOUS REPLACEMENT TECHNIQUE

An experimentally based procedure for solving the phase problem by using several protein crystals containing different heavy atoms.