March 1986

Hcport No. STAN-CS-86-1115
Also numbered KSL-85-41

The Heuristic Refinement Method for Deriving Solution Structures of Proteins

by

Bruce G. Buchanan, Barbara Hayes-Roth, Olivicr Lichtarge,
Russ Altman, James Brinklcy, Micheal Hcwctt, Craig Cornelius,
Brucc Duncan and Olcg Jardctzky

Department of Computer Science

Stanford Univcrsity
Stanford, CA 94305

Knowledge Systems Laboratory
Report No. KSL 85-41

March 1986

The Heuristic Refinement Method for Deriving Solution Structures of Proteins

by

Bruce G. Buchanan, Barbara Hayes-Roth, Olivier Lichtarge, Russ Altman, James Brinkley, Micheal Hewett, Craig Cornelius, Bruce Duncan and Oleg Jardetzky

KNOWLEDGE SYSTEMS LABORATORY
Department of Computer Science
Stanford University
Stanford, California 94305

PREPRINT: Submitted to PNAS 2/86

Classiffcation: biophysics

The Heuristic Refinement Method for Deriving Solution Structures of Proteins

(Nuclear Magnetic Resonance, Structure Refinement,
Artificial Intelligence, Computing)

Bruce G. Buchanan@, Barbara Hayes-Roth+, Olivier Lichtarge+,
Russ Altman*, James Brinkley*, Micheal Hewett.,
Craig Cornelius*, Bruce Duncan*, Oleg Jardetzky+

+Stanford Magnetic Resonance Laboratory
Stanford University Medical Center
Stanford, *CA* 94305

*Knowledge Systems Laboratory
Computer Science Department
Stanford University
Stanford, *CA* 94305

Table of Contents

The Heuristic Refinement Method for Deriving Solution Structures of Proteins

Abstract

A new method is presented for determining structures of proteins in solution. The method uses constraints inferred from analytic data to successively refine both the locations for parts of the structure and the levels of detail for describing those parts. A computer program, called PROTEAN, which encodes this method, has been partially implemented and was used to derive structures for the lac-repressor headpiece from experimental data.

1. Introduction

Elucidation of protein structures to date has been accomplished by X-ray crystallography. This method is limited to proteins which can be crystallized and provides no information on variations in the structure which may occur in solution. As structural information on proteins in solution became available from NMR (1,2), it became apparent that such variations may be important. However, efforts to define procedures which permit a complete independent and rigorous protein structure determination from NMR data alone have encountered significant difficulties (3,4,5).

. We have developed a new method for determining the solution structure of proteins, which relies ***primarily but not exclusively*** on NMR data. The method, called ***heuristic refinement***, proceeds in two steps. First, NMR ***and other*** experimental data obtained from physical measurements on proteins in solution are used to define an ***approximate*** structure which can serve as a plausible starting point for optimization. Second, an optimization procedure is used to further refine that structure. In the first step, we rely entirely on a straightforward interpretation of experimental data. We note that in all cases reported thus far the available set of **experimental** data is not sufficient to completely define the structure, and theoretical constraints must be used to arrive at a specific solution. Because different theoretical constraints may lead to different final structures, we reserve the introduction **of** constraints which cannot be derived from experiment to the second step.

In this report we focus on the first step and illustrate the use of this method on- the lac-repressor headpiece. No crystal structure is available for this protein and two different (but topologically similar) NMR structures have been reported (6.7)

The method we propose and have partly implemented **is a constructive method in which a** complete structure is sequentially constructed from pieces. The definition of tertiary structure follows the definition of the secondary structure, whose inference from NMR data has been extensively described by others **(8-10).** For purposes of this report the secondary structure is taken as a given. The tertiary structure is constructed incrementally in accordance with constraints inferred from experimental data. Some are quantitative: for example, from the observation of a nuclear Overhauser effect (NOE) one can infer plausible ranges of distances between some pairs of protons, and from low angle X-ray scattering experiments one can infer the gross shape and size of the protein. Other constraints are more qualitative: for example, from photo-chemically induced dynamic nuclear polarization (photo-CIDNP) and paramagnetic perturbation experiments one can determine some of the atoms that lie on or near the surface of the molecule. Each of these diverse kinds of constraints weakly restricts the space of plausible conformations: taken together they define the ***range*** of conformations consistent with a given set of experimental data from which optimization methods may start. **The method** is now partly coded in a computer program called PROTEAN, described in sections 3 & 4. Our report is focused on the definition of the tertiary structure since this is the step at which **the** largest uncertainties occur.

2. Sources of Information

The possibility of deriving three-dimensional structures from NMR data rests on the fact that NMR parameters (chemical shifts, coupling constants, relaxation parameters **T1, T2, NOE)** are functions of interatomic distances and (frequently) orientational angles **[14].** Interpreting the measured parameters in terms of structure is difficult because they are also functions of time and of ***a priori*** unknown parameters of the surrounding structure.

The principal though not the only source of interatomic distance information in protein

NMR-spectra are NOE measurements, readily obtained by 2DFT methods [4-7].[*]

Attempts have been made to derive structures from the inherently imprecise data of NMR by imposing additional constraints, such as the requirement that a single structure exists in solution, implied in the use of a distance geometry algorithm [11, 12] or minimizing an energy function which has an equivalent implication [7]. Such algorithmic refinement procedures implicitly introduce a number of uncertainties that we have discussed elsewhere [6]. The proposed method explicitly addresses these uncertainties, including both the imprecision in the constraints inferred from the data and the possibility that the data represent an average reflecting several distinct solution structures.

[*] We note to begin with, however, that *calibration* of NOEs in terms of interatomic distances has no experimental basis. The transfer of magnetization between any two spins I and *k* in an NOE experiment is described by the generalized Bloch equation [10]:

$$dM_i/dt = -\rho_i(M_i - M_i^0) - \Sigma\sigma_{ij}(M_j - M_j^0) + \sigma_{ik}M_k^0$$

where *M* is the magnetization vector, I the observed, k the irradiated and *j* any other interacting spin. The molecular interpretation of the relaxation parameters ρ and σ is given by:

$$\rho_i = Kf_{ij}(\tau)\sum_j 1/r_{ij}^6$$

$$\sigma_{ij} = Kf_{ij}(\tau)/r_{ij}^6$$

where K is a product of atomic constants, r_{ij} the interatomic distance and f_{ij} (l) is a spectral density function for any pair of protons *ij*. A simple relation between the measured magnetization transfer and the internuclear distance therefore exists only if (a) the two spin approximation ● ppliu, i.e. only direct and no indirect magnetization transfer is occuning in the experiment and (b) f_{ij}(l) is accurately known. In protein NMR typically neither is the case [10].

We have recently shown [13] by solving the Bloch equations (1) for a variety of specific structures that even at the shortest experimentally feasible mixing times in an NOE experiment (10-20 msec) indirect magnetization transfer may play a significant role. Thus there is no universally valid relationship between the magnitude of an NOE and internuclear distance. This conclusion also follows from earlier model calculations of Bother-By [5]. In addition, *f(r)* is not known with any accuracy [10].

3. Methods

The problem we are facing is inherently combinatorial: placing the amino acids and atoms of a protein into a 3-dimensional structure in all possible ways that are consistent with known or inferred constraints. The method we use must avoid the combinatorial complexity of this problem. Insofar as interpretations of some of the data yield information that is not easily expressed in precise quantities (e.g., distance ranges, shape or surface information), the method we use must deal with the qualitative nature of the information and not overinterpret the data as precise quantities. For these reasons, heuristic and symbolic reasoning methods used in artificial intelligence (AI) programs are relevant [15].

The heuristic refinement method is based on a reasoning paradigm known as the blackboard model [16]. It was successfully used by Erman, et al., for the interpretation of acoustic speech data [17], by Nii, et al. for the interpretation of acoustic signals emanating from ships at sea [18], and by Terry for the interpretation of electron density maps by the CRYSALIS program [19]. This paradigm encourages (a) the use of different kinds of information, (b) different levels of detail in describing a partial solution, and (c) opportunistic reasoning that selects and focuses on the “best” task to perform next. This method is implemented in the PROTEAN program and is illustrated in the context of the Iac-repressor protein in section 4.’

The kinds of information available to PROTEAN are summarized in’ Table 1. We have initially limited our focus to information inferred from NMR spectra in order to examine the strengths and limits of NMR data without introduction of theoretical assumptions until explicitly called for.

<Table 1 here>†

We believe it is easier to integrate these diverse sources of information in a symbolic

† The data table is included here for purposes of completeness. All data contained in the table have previously been reported by us (27) and by others (20, 28).

inference program‡ and to make explicit -- and change -- the assumptions accompanying them. For example, PROTEAN assumes that an NOE defines a distance range of 2-5 A but this range can be easily changed.

The conventional classification of protein structure into primary, secondary and tertiary provides a useful hierarchy for a stepwise interpretation of NMR and other solution data in structural terms. Methods for protein sequencing are well established. The identification of secondary structures in peptide chains by a combination of NMR measurements of accurate exchange rates and NOEs along the peptide backbone are reasonably reliable [20]. The least determined part of the problem is the three dimensional arrangement of the structured domains.5

The levels of detail considered by PROTEAN are shown in Table 2.

<Table 2 here>

The main reason for introducing layers of abstraction is to reduce the combinatorics of reasoning with many hundreds, or thousands, of individual atoms. The method places larger units, such as alpha helices, in approximate position. Then it has a bounded region in which to place the component parts.

The opportunistic nature of the reasoning is described in Sec. 3.1. Roughly speaking, opportunistic reasoning is what we commonly employ in putting together a jigsaw puzzle. We get the pieces laid out, look for corners and edge pieces, notice uniquely colored or oddly shaped pieces, put together partial "structures" when we can, and build up larger structures from smaller ones. Our focus is not driven by a rigid procedure so much as by opportunities to make progress.

The PROTEAN program itself decides which parts of the structure to work on next and

‡ The PROTEAN program also includes many subprograms that are largely numerical, for example procedures for geometry calculations, but integrates these within a reasoning framework that is non-numeric.

§ We refer to alpha helices and beta sheets as domains.

which information to use. Depending on the available data and the primary and secondary structural features already determined, different problems will be solved in different ways. For example, the program does not start building around an arbitrary domain, nor around the first in the sequence. Instead it reasons about which domain is the "best" starting place, depending on size and the nature of constraints to other parts of the structure. The basic reasoning cycle of the BB1 framework system [16], on which PROTEAN is based, is shown in Figure 1. At any time there may be several actions that are feasible. These actions are the product of knowledge sources (KS's) -- modular and knowledge-intensive pieces of the program that construct part of the puzzle.

<Figure 1 here>

3.1. Method of Inference: Heuristic Refinement of Accessible Volumes and Descriptive Detail

Our basic strategy is to examine constraints one by one (or in subsets) and to define the region of space that two structured domains can occupy with respect to each other, still satisfying the given constraints. This region can be specified by a table of the six parameters (three for position and three for orientation) that describe the three-dimensional relationship of two solid objects, and it can be progressively restricted by the introduction of additional compatible constraints. As discussed more fully below, finding that two or more constraints are *incompatible,* (i.e., that the accessible volumes and/or orientations defined by the constraints do not overlap) allows the conclusion that the protein does *not* exist in a single conformation.

In addition to refining the volumes accessible to parts of the structure, PROTEAN also refines the structural description by describing the protein at successively finer levels of detail, as shown in Table 2. In our current implementation, we use only one level of detail, the solid level, in which alpha helices, beta strands, and random coils are represented as single units. By reasoning with aggregates of atoms in this way, the program is able to avoid the combinatorial explosion of possible positions of all atoms together, but still leads to a definition of the topology.

Just as with solving jigsaw puzzles, the procedure followed for any one problem depends (opportunistically) on the nature of the problem. The general strategy is to iterate over the following steps:

- (a) choose a level of detail, and determine the component parts of the protein at that level;

- (b) select which part(s) to consider as a partial solution:

- (c) position the component parts of the partial solution according to constraints between them; and

- (d) combine partial solutions.

Initially, the program defines the primary and secondary structures, specifies constraints, and chooses one domain (e.g., a helix) as an "anchor". For a helix as anchor, a convenient choice of coordinate system is to define the axis of the helix as the z-axis of a right-handed Cartesian coordinate system and the x-axis as a perpendicular axis passing through the nucleus of the first (N-terminal) alpha-carbon in the helix. The program then selects a second structured domain to position relative to the first by considering which of the unpositioned domains has the most and the strongest constraints to the initial anchor. Then the second domain is positioned relative to the first by sweeping out the accessible .volume in which the second domain can be located and still satisfy each constraint between them. This volume represents the limits of our knowledge about the structure defined by a subset of the experimental constraints.

After an anchor is chosen and a second domain is positioned relative to it, the next step is to choose either another constraint or a new domain to work on, or to choose to describe part of the structure at a finer level of detail. With the introduction of each constraint, the allowed locations of domains relative to one another are further restricted. With the introduction of each new domain, the partial structure is expanded. And with increased detail in the description of any part, locations of individual atoms are more closely determined.

When a new domain is selected, it is positioned relative to the initial anchor and to other domains with respect to constraints between them in an order decided by applicable knowledge sources. Then its allowed positions are further refined by considering the intersection of the consequences of all its relative positions together. These steps are applied until the list of structured domains is exhausted. This is followed by one or more steps of satisfying the constraints (if any are available experimentally) between any unstructured ("random coil") segments and the domains defined by secondary structure.

The knowledge **sources that** place parts of the protein into the evolving solution, or that refine the allowed locations of a piece based on a new constraint, call on highly specialized procedures for geometric reasoning. In principle, the geometry calculations simply sample all of space at some defined resolution, retaining all sample locations that satisfy the constraints. In practice, intelligent selection of which regions of space to sample avoids the computational impossibility of defining a continuous volume by sampling an infinity of locations. Nevertheless, the program may define the allowed locations of a helix (with respect to another domain) by listing a thousand or more locations.

Because we save only the locations that satisfy all the constraints considered so far, at every point in the procedure we have a partial definition of a structure that is accurate with respect to constraints considered, even if lacking in precision. If the accessible volume of a domain is reduced to nil with the introduction of new constraints, we can conclude that the constraints cannot be satisfied simultaneously, and therefore that no single structure exists. This may be the case, for example, in a protein structure fluctuating between two or more conformations. Otherwise, we know that the constraints can be satisfied simultaneously and that a single structure ***may*** exist. We then have a choice: (1) to ***assume*** that the constraints ***are*** satisfied simultaneously, and thus start with the reduced (intersected) family of positions when exploring new constraints: or (2) to acknowledge that there may be important conformations for which all the constraints are ***not*** satisfied at the same time, and so work with the full set of positions implied by ***each constraint,*** reasoning with combinations of them to define plausible families of conformations. The latter procedure is considerably more time consuming, but is more cautious. PROTEAN allows either choice.

It is worth noting that long range NOE constraints contain information on the possible geometry of macromolecular configurations, but not on their **energetics.** Thus, the fact that **a** combination of constraints is geometrically possible provides no information as to whether it is energetically feasible. This is an important limitation of the information content of a spectroscopic measurement and requires great caution in the use of the very tempting assumption that any two constraints which **can** be satisfied simultaneously indeed **are** satisfied simultaneously. **A** much finer discrimination between constraints that can and cannot be satisfied simultaneously can in principle be made by the refinement procedure based on **the** solution of **Bloch** equations described below **[Sec.4.2].**

3.2. The PROTEAN Program

PROTEAN is still under construction and currently reasons at the solid level of detail using NOE constraints in addition to general knowledge about protein chemistry. The program currently assumes that a biochemist will select which constraints to relax when no structures can be found which satisfy all constraints simultaneously. We are extending the program to reason at the superatomic and atomic levels, to reason with other kinds of constraints, and to reason about constraints that are not satisfiable simultaneously, as outlined above. Some preliminary results are shown in the next section.

4. Results and Discussion

4.1. Correctness of the Method

The method is cautious in the following sense: (a) it introduces no theoretical assumptions about protein structure that are not inherent in the experimental methods used, (b) it uses broad ranges for distances inferred from NOEs, (c) it includes ***only*** locations that are consistent with the stated constraints, (d) it includes as fine-grained a sample of ***all*** allowed locations as specified, subject to limitations on computer time and storage. Opportunistic reasoning reduces the time required to solve a structure (at some resolution) by carefully selecting the order in which constraints are introduced. But the result ‘of the refinement is insensitive to the order¶. Because all the stated constraints are used, and they are used, essentially, to rule out locations of component parts, the space of locations will be as small as allowed by the constraints and will contain all allowed structures. For these reasons, we believe the method produces an accurate family of the structures within the limited information content of a set of constraints.

We have tested our implementation of the method against the known crystal structure of myoglobin. Starting with atomic coordinates as defined in [21], we first selected half the molecule (four helices) to work with to simplify testing. We then calculated which pairs of hydrogen atoms were within 3.25 A of one another, as a way of defining atom pairs that almost certainly give rise to NOEs (placing the atoms within 2-6 A from each other) if an NMR spectrum were to be taken of the crystal structure of myoglobin. We then used these computed NOE distance ranges, plus the primary and secondary structure, as the input to

¶ I.e., within small error bounds introduced by the resolution of the sampling of positions.

PROTEAN. PROTEAN was run at the solid level of detail (See Table **2),** and produced a family of structures whose members were consistent with the constraints. The basic question to be answered was whether PROTEAN's definition of structure is accurate. That is, does the known crystal structure lie within the family defined by **PROTEAN?**

The experiment is reported in detail in **[22].** Briefly summarized, the answer it provides is positive: PROTEAN successfully defines a family of structures that includes the known crystal structure of **myoglobin.**

4.2. Construction of the lac-repressor structure at the solid levei of abstraction

The lac-repressor headpiece is a protein with fifty-one amino acids, **whose complete** solution structure is not known and which is the subject of current structural studies. **[6, 7]** We illustrate **PROTEAN's** problem solving behavior with its construction of the lac-repressor structure at the solid level of detail.

Table 3 shows the primary structure, secondary structure, and distance constraints inferred from **NOEs||** that are used as input data for PROTEAN. More general protein information, such as **peptide** geometry and Van der Waals radii, is also available to the program.

<Table 3 here>

When **PROTEAN** activates the **knowledge source ACTIVATE-ANCHORSPACE,** it selects the helix with the greatest constraining power, Helix-l **(6-14)** and places it in the initial coordinate system as the anchor. The system then chooses the most highly constrained domain Helix-3 **(35-45),** to add to **this** coordinate system. For each relevant distance constraint between Helix-l and Helix-3, the **knowledge** source EXPRESS-NOE-CONSTRAINT specifies the region of space for Helix-3 that is compatible with **the** constraint. In this case, the

|| We have shown elsewhere [5, 8, 20] that for structures of this size both *direct* and *indirect* long range NOEs obtained at mixing times of 100 msec or less imply an upper distance limit of about 6 A and have therefore included both in our data base. (At the solid level, we translate an NOE constraint into an rpproximrte distance of 15 A between points on the solids corresponding to the C_α's of the appropriate peptides.) The distinction becomes important only in the refinement procedure.

intersection of these regions is taken to define the net accessible volume of Helix-3 with **respect** to Helix-l. The. knowledge source ANCHOR-HELIX performs this intersection in order to specify the complete **set** of helix locations for which all distance constraints can be satisfied simultaneously. Similarly, the knowledge source ANCHOR-COIL can specify the corresponding set of locations for the random coil segments.

Having placed Helix-3 relative to Helix-l, ANCHOR-HELIX then places Helix-2 relative to the anchor, Helix-l. Constraints between these two anchorees, Helix-2 and Helix-3, (but not involving the initial anchor directly) can be used to further restrict their accessible volumes, and this is done by the knowledge source YOKE-STRUCTURES.

The result defines the main topological features of the molecule, as shown in Fig. 2, and the accessible volumes within which elements of the structure remain uncertain. These volumes indicate the extent to which the structure can be specified from the existing solution data at this level of detail. It is seen from the figure that the accessible volumes are in this case sufficiently large to preclude a unique definition of a single structure, but also sufficiently small to recognize major structural relationships.

<Figure 2 here>

4.3. Discussion

The accessible volumes defined for each of the structured domains and random coils are, strictly speaking, measures of the uncertainty of spectroscopic data. This uncertainty has two principal sources: insufficient data and degrees of internal **motional** freedom. The existence of mutually exclusive structural constraints can be taken as definite evidence that internal motion is occurring. This is not found to be the case with the lac-repressor data. In the absence of truly incompatible constraints the distinction between insufficiency of the data and internal mobility is difficult and can only be made on additional experimental **grounds. For** example, if the relaxation of nuclei on domain A by nuclei on domain B does not follow the pattern that **can** be predicted from any single structure which can be instantiated within the **accessible volume, it is possible to infer that** internal motion must be occurring. Obviously, **extensive additional experimental evidence** is required to substantiate such conclusions.

Several options exist for specifying the structure with greater precision. They fall into two classes: (a) refinement. by theoretical constraints and (b) refinement by additional experimental data. Into the first category fall the use of a distance geometry algorithm [11], optimization procedures [24], energy minimization [13] and relaxation of equations of motion in a molecular dynamics calculation [25]. Common to all of these procedures is the untestable *a priori* assumption that all observed constraints are satisfied simultaneously and hence that a single structure corresponding to a global energy minimum exists? An example of a structure arrived at by such a procedure -- an optimization algorithm developed at Stanford [24] -- is shown in Fig. 3. The structure which corresponds to the minimum of an error function falls well within the accessible volumes defined by PROTEAN. It bears a strong similarity to but also shows significant differences from the structure arrived at by an alternative theoretical refinement procedure used by Kaptein and coworkers [16]. The open issue is therefore whether the theoretically based refinement procedures can give a unique and accurate representation of the "real" structure.

<Figure 3 here>

Our clear preference is for refinement procedures based on additional experimental data. We have recently used a refinement procedure based on the solution of the Bloch equations (1), taking all alternative pathways of magnetization transfer between a given pair of spins into account [26]. This procedure has been applied to specify the structure of the trp operator DNA. Although it has not yet been applied to the lac-repressor structure, it is briefly outlined here for completeness. The procedure rquires that the complete time course of magnetization transfer be experimentally measured for each observed NOE. An instance of a structure within the accessible volumes defined by PROTEAN is then selected and for each NOE constraint the time course of magnetization transfer is calculated by solving the Bloch equations taking into account alternative pathways suggested by the local constellation of spins in the selected

**A family of structures obviously clustered about a single average, as is often obtained using the distance geometry algorithm [11], defines the uncertainty limits of a single structure. We consider it preferable to speak of multiple structures only when there is a clearly bimodal distribution of such clusters.

structure. The experimental and the calculated curves are then matched and if necessary the structure is modified (using a gradient search routine) until the observed and calculated NOE buildup curves are brought into agreement. All structures which give a correct prediction of the experimentally observed magnetization transfer must be accepted as "real".

This refinement procedure shares with the theoretically based procedures the need for a *'reasonable" starting structure. The family of plausible structures is given by the output of the PROTEAN program. The explicit constraints given to PROTEAN at present do not distinguish any instances as more plausible than others, but further data or assumptions could provide a single instance as a "very plausible" starting place for theoretically based procedures. Thus, it is possible to couple PROTEAN and mathematical refinement procedures. Even without these additional procedures, however, the method of heuristic refinement outlined here is a means of determining the families of positions consistent with any limited set of experimental constraints.

5. Conclusions

The key features of the problem -- (1) the fact that a one-to-one correspondence between experimentally measured and molecular structural parameters does not hold' for solution methods as it does in crystallography, (2) the need to use constraints. that are more easily expressed symbolically than numerically, (3) the availability of different kinds of constraints, (4) the incompleteness and irreducible uncertainty of the data -- make it desirable to follow more than one option in the analysis of the data and suggest that **AI** methods are appropriate.

One of the key features in the design of the heuristic refinement method is flexibility in incorporating different kinds of information. Preliminary results from using only constraints inferred from experimental NMR data indicate that additional kinds of constraints will be necessary to refine the allowed positions of **helices,** beta strands, and coils. These may come from additional experimental data and from theoretical considerations such as energy minimization. **A** second key feature is that the method does not overinterpret the data, but defines the ***family*** of structures compatible with the **data.**

Acknowledgments

We thank **A.N.** Lane and J.F. Lefevre for helpful discussions: F. **Frayman** for the coordinates of the structure in Fig. **3;** Jeff Harvey, Alan **Garvey** and John Brugge for programming assistance; Grace Smith and **Anthea Waleson** for manuscript preparation.

This work was funded in part by: NIH Grant #GM07365, NIH Grant #RR-00785, NIH Grant #RR-00711, NSF Grant #DMB84-02348, NASA/Ames Grant #NCC 2-274, Boeing Grant #W266875, DARPA Contract #N00039-83-C-0136.

REFERENCES

1. Roberts, G.C.K. and Jardetzky, 0. (1970) Adv. ***in Protein Chem.*** **24:447-545.**

2. Wutrich, K. (1976) "NMR in Biological Research: **Peptides** and Proteins", North Holland, Amsterdam.

3. Jardetzky, 0. and Roberts, G.C.K. (1981) "NMR in Molecular Biology", Academic Press, New York.

4. Wagner, G. and Wutrich K. (1979) ***J. Mag. Res.*** **33:675-679.**

5. Bothner-By, A.A. and Noggle, J.H. (1979) ***J. Am. Chem. Soc.*** 101, 5162-5170.

6. Jardetzky, 0. (1984) IN: Progress in Bioorganic Chemistry and Molecular Biology, Yu. A. Ovchinnikov (ed), Elsevier Science Publishers B.V., Amsterdam.

7. Kaptein, **R.,** Zuiderweg, E.R.P., **Scheek,** R.M. and Boelens, R. (1985) ***J.*** Mot. ***Biol.*** 182, 179482.

8. Wutrich, K., Wider, **G.,** Wagner, G. and Braun, W. (1982) J. Mot. ***Blot.*** **155:311-319.**

9. Wemmer, D. and Kallenbach, N.R. (1983) ***Biochemistry*** **22:1901-1906.**

10. Kalk, A. and Berendsen H. J. (1976) ***J.*** Mug. ***Res.*** **24:343-366.**

11. Havel, T.F., Crippen, G.M. and **Kuntz,** I.D. (1979) ***Biopolymers*** **18:73-81.**

12. Braun, W., Bosch, C., Brown, L.R., Go, N. and Wutrich, K. (1981) ***Biochim. Biophys. Acta*** 667, 377.

13. Lane, A.N., **Lefevre,** J.F. and Jardetzky, 0. unpublished data.

14. Jardetzky, 0. (1964) Adv. Chem. Phys. VII, J. Duchesne (ed.), Interscience, NY, 499-532.

15. Rich, E. (1983) ***Artificial Intelligence New*** York: McGraw-Hill.

16. Hayes-Roth, B. (1985) ***Artificial Intelligence.*** **26:251-321.**

17. Erman, L.D., Hayes-Roth, F., Lesser, V.R., and Reddy, D.R. (1980) ***Computing Surveys*** **12:213-253.**

18. Nii, H. P., Feigenbaum, E.A., Anton, JJ. and Rockmore, **A.J.** (1982) AI ***Magazine*** **3:23-35.**

19. Terry, A. (1983). Ph.D. Thesis (Information and Computer Science), University of California, Irvine.

20. Zuiderweg, **E.R.P.,** Kaptein, R. and Wuthrich, K. (1983) ***Proc. Natl. Acad. Sci. 80,* 5837-5841.**

21. Watson, H.C., (1969) ***Prog. Stereochem. 4, 299-333***

22. Lichtarge, **O.,** Buchanan, B., Cornelius, C., Brinkley, J., Jardetzky, 0. (1986) ***Knowledge Systems Laboratory Technical Report KSL-86-12,*** Stanford University.

23. Lane, A.N. and Jardetzky, 0. unpublished results.

24. **Frayman, F.** (1985) **"PROTO."** Ph.D. Thesis (Computer Science), Northwestern University.

25. Karplus, M. and **McCammon,** J.A. (1983) ***Annu. Rev. Biochem.* 18:927-942.**

26. Lane, A.N., **Lefevre,** J.F. and Jardetzky, 0. unpublished results.

27. Ribeiro, **A.A.,** Wemmer, D., Bray, R.P. and Jardetzky, 0. (1981) ***Biochem. Blophys. Res. Comm.* 99:668-674.**

28. Zuiderweg, E.R.P., Billeter, M., **Boelens,** R., **Scheek,** R.M., Wuthrich, **K.**· and Kaptein, R. (1984) ***FEBS Lett.* 179, 243-247.**

List of Figures

I. **General** knowledge about protein chemistry, e.g.,

composition of amino acids
peptide bond geometry
standard conformations of alpha helices
van der Waals radii

II. Global constraints about the structure inferred from experimental data, e.g.,

overall shape
density
surface atoms (from solvent accessibility)

III. Local constraints about relative positions of atoms inferred from experimental data, e.g.,

approximate distances between some pairs of atoms (from NOEs)

Table 1. Some types of information available to PROTEAN. Global constraints are not currently used.

Molecular level	Consider the entire molecule as a single unit.
Solid level	Consider alpha helices, beta strands, and random coils as separate geometrical solids. An alpha helix, for example, is represented as a cylinder.
Superatomic level	Explicitly represent some or all covalent bonds around which rotation is possible and consider the structural units between them as single units called superatoms. Alanine, for example, may be represented as three superatoms: the peptide group, the alpha-CH, and the CH_3 group.
Atomic level	Consider the individual atoms in the structure. (This target level is not achievable for all parts of the structure using only NMR data.)

Table 2. Levels of detail to be used in PROTEAN for reasoning about and describing protein structures. At present, only the solid level is used.

Primary Structure:

MET1 LYS2 PRO3 VAL4 THR5 LEU6 TYR7 ASP8 VAL9 ALA10 GLU11 TYR12 ALA13 GLY14 VAL15 SER16 TYR17 GLN18 THR19 VAL20 SER23 ARG22 VAL23 VAL24 ASN25 GLN26 ALA27 SER28 HIS29 VAL30 SER31 ALA32 LYS33 THR34 ARG35 GLU36 LYS37 VAL38 GLU39 ALA40 ALA41 MET42 ALA43 GLU44 LEU45 ASN46 TYR47 ILE48 PRO49 ASN50 ARG51

Secondary Structure:

Coil-l: MET1-THR 5
Helix-l: LEU6-GLY 14
Coil-2: VAL15-SER16
Helix-2: TYR17-ASN25
Coil-3: GLN26-ARG35
Helix-3: GLU36-LEU45
Coil-4: ASN46-ARG51

Observed NOES between amino acid pairs:

VAL4, TYR17
VAL4, LEU45
VAL4, TYR47
THR5, TYR47
LEU6, TYR17
LEU6, VAL24
LEU6, MET42
LEU6, TYR47
TYR7, TYR17
ASP8, LEU45
VAL9, MET42
VAL9, LEU45
VAL9, TYR47
ALA10, VAL20
ALA10, TYR17
TYR12, ALA32
TYR12, ALA41
TYR12, MET42
TYR12, GLU44
TYR12, LEU45
ALA13, VAL38
ALA13, VAL41
VAL15, TYR47
TYR17, MET42
VAL20, VAL38
VAL24, TYR47
VAL30, MET42
MET42, TYR47

Table 3. **Input data for** PROTEAN's **determination of the partial structure of the lac-repressor headpiece.**

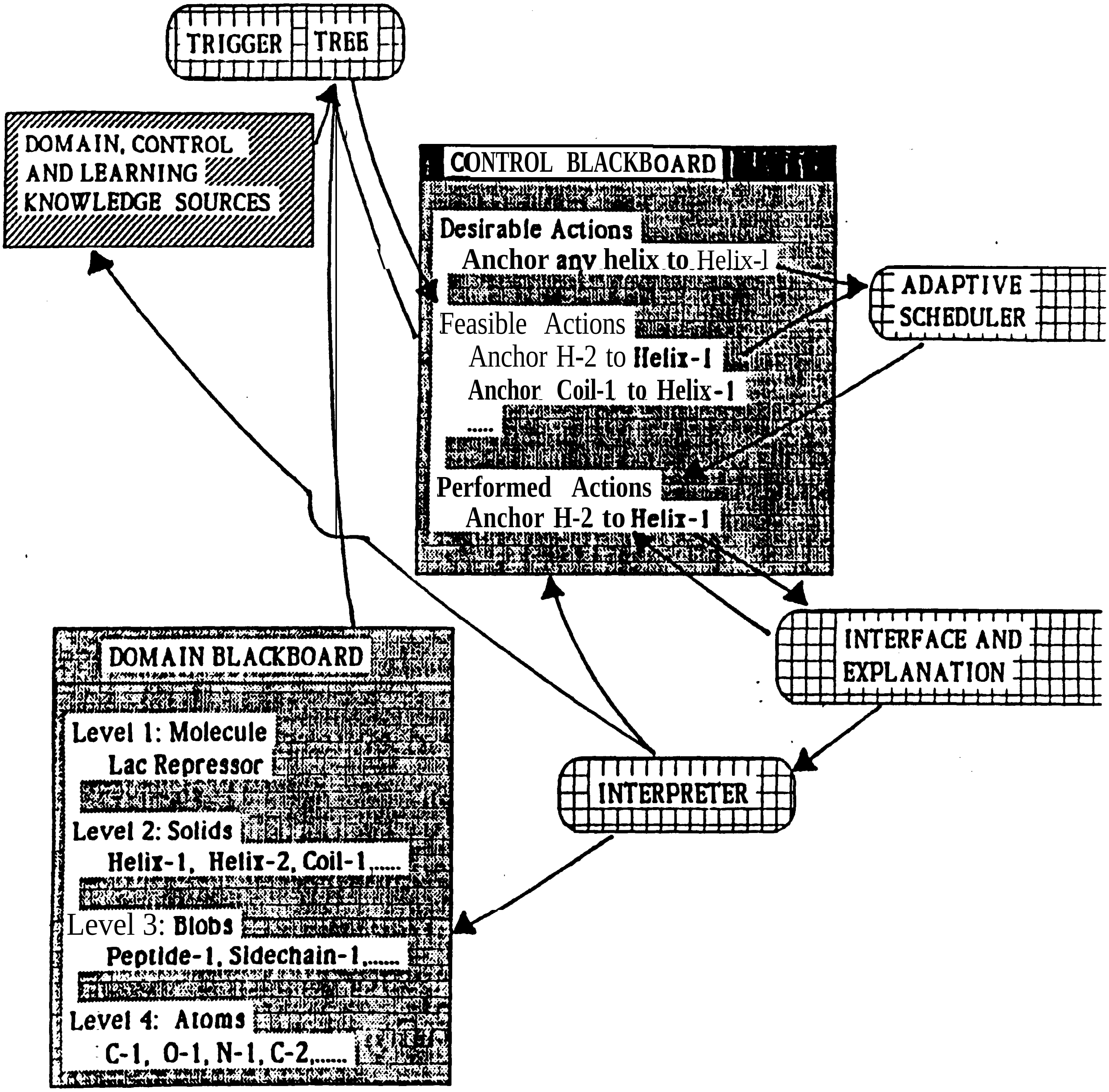
TRIGGER TREE
DOMAIN, CONTROL AND LEARNING KNOWLEDGE SOURCES
CONTROL BLACKBOARD
Desirable Actions
Anchor any helix to Helix-1
Feasible Actions
Anchor H-2 to Helix-1
Anchor Coil-1 to Helix-1
......
Performed Actions
Anchor H-2 to Helix-1
ADAPTIVE SCHEDULER
INTERFACE AND EXPLANATION
INTERPRETER
DOMAIN BLACKBOARD
Level 1: Molecule
Lac Repressor
Level 2: Solids
Helix-1, Helix-2, Coil-1,......
Level 3: Blobs
Peptide-1, Sidechain-1,......
Level 4: Atoms
C-1, O-1, N-1, C-2,......

FIGURE 1

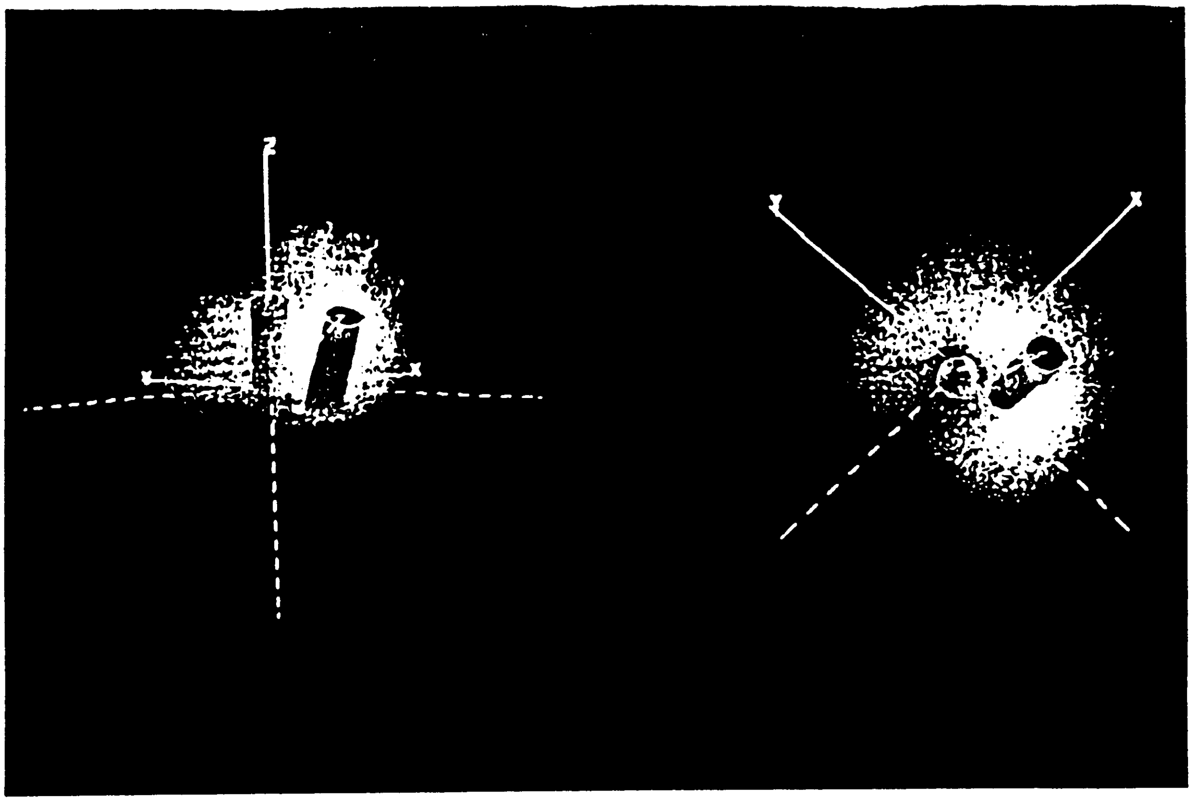

FIGURE 2

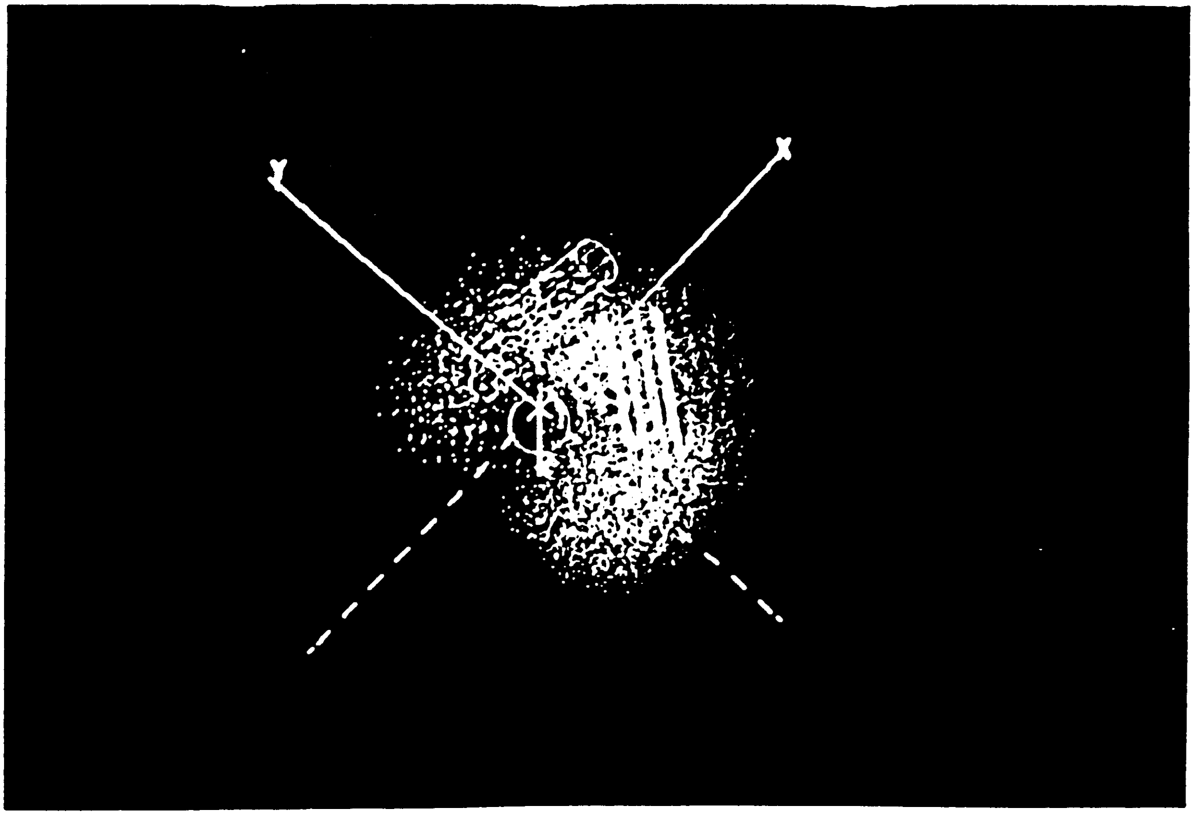

FIGURE 3