# A knowledge-based architecture for protein sequence analysis and structure prediction

**Dominic A. Clark, Geoffrey J. Barton, and Christopher J. Rawlings**

*Biomedical Computing Unit, Imperial Cancer Research Fund, Lincoln's Inn Fields, London, England*

*Methods for analyzing the amino-acid sequence of a protein for the purposes of predicting its three-dimensional structure were systematically analyzed using knowledge engineering techniques. The resulting entities (data) and relations (processing methods and constraints) have been represented within a generalized dependency network consisting of 29 nodes and over 100 links. It is argued that such a representation meets the requirements of knowledge-based systems in molecular biology. This network is used as the architecture for a prototype knowledge-based system that simulates logically the processes used in protein structure prediction. Although developed specifically for applications in protein structure prediction, the network architecture provides a strategy for tackling the general problem of orchestrating and integrating the diverse sources of knowledge that are characteristic of many areas of science.*

*Keywords: protein sequence analysis, knowledge-based systems, knowledge representation, logic programming*

A long-term goal of molecular biology is the prediction of protein function from protein sequence. It is generally agreed that an important aspect of this function prediction is structure prediction. Techniques for protein structure prediction fall into two classes: (1) methods aiming to predict from sequence information alone using theoretical models such as molecular dynamics and energy minimization; and (2) empirical methods, which attempt to combine sequence data with other potentially relevant information such as the known structure of a homologous protein. Of the methods that combine information with the protein sequence in predicting tertiary structure, the most successful developments have been in the field of protein model building.[1] Recent developments have sought to automate aspects of this procedure.[2,3] Although prediction based on model building from a homologous 3D structure is likely to yield the most reliable structures, below 30% sequence identity it is essential to seek other (biophysical/biochemical) evidence for the model-built structure. Furthermore, since model building techniques demand knowledge of the 3D structure of a homologous protein, they go only some way to redressing the information gap between the number of known sequences and structures. Both to corroborate model-built structures and to predict the structures of proteins whose sequence identity with a known 3D structure falls below 30%, there is a longer-term need to investigate methods to systematically automate the use of diverse data in prediction.

In this paper we introduce the concept of knowledge-based systems as a method for coherently integrating data analysis techniques relevant to protein structure prediction. A study is then described that utilizes techniques from knowledge engineering[4,5] to analyze the problem of protein structure prediction. The relative merits of flowcharts and networks as knowledge-based architectures are discussed. A prototype system based upon a network model that can assist scientists in the analysis and interpretation of protein sequence and other data is described and the general problems of building large systems that integrate diverse knowledge are discussed.

## KNOWLEDGE ENGINEERING

Knowledge engineering is the process of specifying the descriptive and strategic knowledge necessary to perform a task,[5] usually derived from human experts or textual documentation. Techniques for predicting protein structures on the basis of model building are well documented.[2,3] In this study, we sought to produce a broader characterization of knowledge (both descriptive and strategic) relevant to predicting protein structure from sequence by extending the focus beyond the analysis of model-built predictions to published papers that predicted protein structure without explicit use of an homologous protein of known structure. The set of papers analyzed predicted the structure of interferon,[6] interleukin-2,[7] human growth hormone (henceforth HGH),[8] $\alpha$ subunit of tryptophan synthase,[9,10] human epidermal growth factor receptor,[11] and cation transporting ATPases.[12] These papers typically employed the protein sequence in conjunction with biophysical/biochemical data, and other infor-

Address reprint requests to Dr. Clark at the Biomedical Computing Unit, Imperial Cancer Research Fund, P.O. Box 123, Lincoln's Inn Fields, London, WC2 3PX, UK.
Received 5 December 1989; accepted 9 January 1990

mation derived from known 3D structures (e.g., methods of secondary structure prediction, known folding and packing constraints, techniques of sequence alignment) to arrive at a *plausible* tertiary structure prediction.

Analysis of the prediction papers involved identifying the logical organization of analyses, experiments, and arguments presented (typically the order in which journal papers are presented). In general, the strategy employed by most sets of authors involved attempting to produce the most consistent interpretation of the broadest range of data. The analysis of one prediction paper appears in Table 1. Details of some other papers appear in the appendixes.

In general, the sequence of events presented in published papers can be more of a post hoc rationalization than a chronological description. However, from the perspective of constructing a knowledge-based system, it is the logical manner in which information is combined to produce the "argument" for the proposed structure that is important rather than the chronology of the information processing.

## Knowledge-based systems

In contrast to traditional computer software, knowledge-based systems can be characterized by the separation and explicit representation of descriptive knowledge (facts, assumptions, and hypotheses) and strategic knowledge (how to use the descriptive knowledge to solve problems), combined by logical inference (see Rawlings et al.[13]). Under this characterization the model-building programs described by Sutcliffe et al.[2,3] are not knowledge-based systems. The separation of descriptive and strategic knowledge makes knowledge-based systems flexible, modular, transparent, and robust, and allows reasoning about high-level relationships and control. For the scientific community, the usefulness of knowledge-based support systems comes from the following:

1. The opportunity to *integrate diverse sources of information* that are relevant to a problem
2. The ability to be *exhaustive* in determining the ramifications of hypotheses and inferences
3. *Neutrality* in terms of the way in which hypotheses are assessed
4. *Research coordination,* by reduction of memory load and maintenance of sources and justification for each inference and hypothesis
5. *Experimentation* with hypotheses and lines of reasoning
6. *Consistency and truth maintenance* of sets of facts and hypotheses

These are all essential in any scientific task involving a wide range of heterogeneous data or knowledge, especially when there is uncertainty. An illustration of the importance of consistency maintenance in protein structure prediction is provided by Hurle et al.[9] who assigned the $\alpha$ subunit of tryptophan synthase to the $\alpha/\beta$ structural class because its CD spectra and secondary structure composition were most consistent with a reference $\alpha/\beta$ barrel structure. Subsequently, however, the barrel structure (later corroborated by x-ray data) was inconsistently rejected in favor of an $\alpha/\beta$ sheet topology. In this prediction, the inconsistency is clear. In general, however, inconsistencies can be difficult to spot within a large quantity of experimental, analytic, and conjectural data. A knowledge-based approach to the orchestration of protein sequence analysis is beneficial because inconsistencies become formally identifiable through their declarative representation.

## Knowledge representation for protein sequence analysis and structure prediction

To successfully integrate knowledge, databases, existing software, and data (biochemical, biophysical, etc.) in a knowledge-based application in a field such as molecular biology, which is both diverse and rapidly changing, a formal architecture is required.

***Flowcharts*** To date, the most practical description of the processes and decisions involved in the analysis of sequence data for predicting the structure of a protein by model building is Taylor's flowchart of "possible paths to follow in the prediction of structure" (Figure 1),[14] although other strategic models have also been proposed.[15] In Figure 1, rectangular boxes correspond to processes (such as database search, secondary structure prediction, hydropathy profiling, alignment, and model building), diamonds are decision points, and arrows indicate flow of control. Rawlings has demonstrated how this strategic model can be represented as a set of rules in the knowledge representation language, PROPS2.[16–18] However, although Taylor's flowchart provides one plausible strategy for predicting protein structure with certain types of information, we argue that flowcharts cannot be considered as the basis for a general knowledge-based architecture for protein sequence analysis and structure prediction.

By definition, flowcharts embody a model of control in which the order of process execution is specified through parameters assessed by decision points. This is appropriate for the formal specification of algorithms, but is antithetic to a practical knowledge-based architecture that might assist protein sequence analysis and structure prediction for the following reasons:

1. Many of the processes involved (e.g., hydropathy analysis and secondary structure prediction) are potentially and meaningfully applicable at many stages in the prediction process. So, for example, in contrast to Figure 1, there are clearly many contexts in which hydropathy analysis might be usefully applied before secondary structure prediction (e.g., to assist in the identification of transmembrane regions).
2. Flowcharts demand the enumeration of all potential outcomes with strategic implications from each process and the most appropriate course of action for each case. As the number of processes in the system increases this becomes impractical, a difficulty encountered by the authors during initial attempts to expand the scope of Figure 1.[19]
3. Practically, flowcharts do not facilitate incremental extension because of the requirement to revise existing and perhaps complex dependencies.
4. Flowcharts cannot be used intelligently unless the available information concords with that expected by the flowchart. For example, a flowchart that expects the user to have one protein sequence and no other data may be irrelevant when a large set of biochemical and biophysical data or a family of aligned sequences are already available.

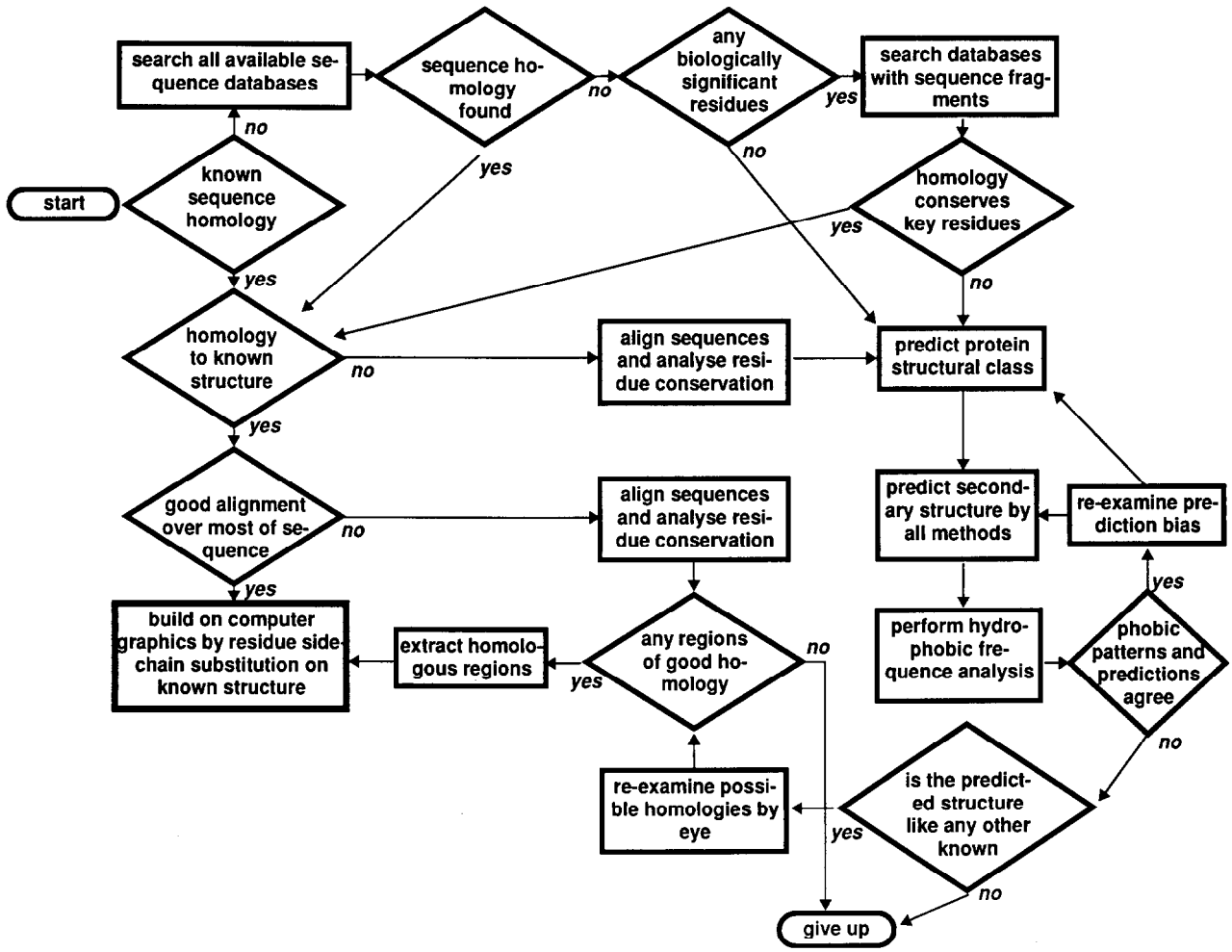


*Figure 1. Possible paths in the prediction of protein structure (Ref. 14)*

A practical knowledge-based architecture for protein structure prediction should therefore satisfy the following requirements:

- *Tractability*. That it is tractable to coherently represent large amounts of diverse knowledge within the same framework.
- *Flexibility*. That strategic knowledge is separated from descriptive knowledge to permit a less constrained representation of the temporal relationships among the relevant entities (objects/data) and processes to allow flexibility, user intervention, and experimentation in control regimes.
- *Modularity*. That the representation be sufficiently modular to permit incremental extension. These requirements are more fully met by networks than flowcharts.

***Networks*** A network is a set of nodes and links that represent entities and relations. The particular network described here draws some inspiration from blackboard models of problem solving,[20,21,22] which have been employed on problems analogous to protein structure prediction, such as speech recognition,[23] and in areas of biology, such as interpretation of 2D NMR data[24] and low-resolution electron density maps.[25,26] The differences between network representations and flowcharts such as Figure 1 are summarized below:

- *Representation*. In networks, nodes represent entities and links represent processes, whereas in flowcharts nodes represent processes or decision points and links represent the flow of control.
- *Modularity*. The local nature of dependencies between nodes and links in networks makes networks more modular than flowcharts and therefore better suited to incremental development.
- *Initiation*. Whereas flowcharts have defined start and end points, no such constraint exists for networks, with the implication that the user is able to start with any collection of knowledge.
- *Knowledge*. Whereas flowcharts are built around strategic knowledge, networks are primarily descriptive, but permit the superimposition of strategic knowledge. This separation of strategic and descriptive knowledge permits the specification of control regimes of arbitrary complexity and/or generality and simultaneously allows flexibility in the order in which processes are executed. Therefore, any high-level strategy (e.g., as embodied in Figure 1) can be superimposed on a network.

- *Functionality*. Network models can be used as the basis of many decision support functions such as browsing and critiquing user plans,[27] while flowcharts models are limited to giving strategic advice for a limited set of situations.

Networks therefore have the tractability, flexibility, and modularity required for a practical knowledge-based architecture for protein sequence analysis and structure prediction.

## THE KNOWLEDGE BASE

Analysis of prediction papers revealed many information sources in addition to those in Figure 1. These included information from biochemical and biophysical assays such as proteolytic cleavage[10] (Appendix 1), chemical cross-linking[10] (Appendix 1), mutagenesis data[10] (Appendix 1), NMR data, circular dichroism spectra[8,9] (Appendix 2), and disulfide linkage[8] (Table 1). These analyses also highlighted the importance of topological reasoning, functional argumentation,[12] and significant patterns, regions, and residues.

**Table 1. Summary of analysis of stages in the prediction of structure of HGH**

| Stage | Description |
|---|---|
| 1 | Initial information: Protein sequence, CD spectra, S-S bridges, Protein ID, Functional Class, Secondary structure composition from CD. |
| 2 | The structural class was predicted from the secondary structure composition suggested by CD. This was 45–50% $\alpha$ and no $\beta$ structure so the protein was assigned to the all-$\alpha$ structural class. |
| 3 | Secondary structure prediction was performed using (a) Cohen turn prediction (all-$\alpha$), (b) hydrophobic "diamonds" to suggest helix positions, (c) "delimit" methods to suggest the ends of helical regions. One unique secondary structure was carried forward consisting of four core helices which accounted for 80–85% of the secondary structure expected from CD. |
| 4 | The Cohen helix packing algorithm was applied to the secondary structure elements and gave 543 different folds (tertiary structures for the 4 core helices). |
| 5 | Constraints from S-S bridge connectivities were introduced, specifically {$Cys_{53}$, $Cys_{165}$} and {$Cys_{182}$, $Cys_{189}$} leading to a reduced list of folds of 186. |
| 6 | Further constraints added were that the structure should have a low surface area to volume ratio and that long loop excursions were not normally allowed. This led to 67 remaining folds. |
| 7 | Structural class specific folding constraints applied were that only right-handed 4-helical bundles had been previously observed in x-ray structures. This constraint cut the list to 5 similar structures. |
| 8 | Consistency check against crystal structure. Not consistent because the observed structure actually has a long loop extension and is a left-handed four helical bundle. The constraints applied at stages 6 and 7 were too limiting. |

*Source*. Reference 8.

The network model that represents the various types of knowledge and data is presented in Figures 2 and 3. Figure 2 shows the complete set of entities and knowledge identified during knowledge engineering. Figure 3 is a subgraph of Figure 2 that pertains to the prediction of the 3D structure of HGH (Table 1). In these figures, nodes represent entities and links (arcs) represent relations between entities, which are either processes (the application of software, knowledge-based inference, biochemical/biophysical assays) or constraints (requirements for consistency).

### Entities

The entities in Figure 2 can be grouped informally into five categories (Table 2). These are biological substance, structural description, classifications and identifiers, results of biophysical and biochemical assays, and results of database queries, sequence analyses, and other software. This grouping is informal since it is the links rather than the nodes that correspond to processes, and values for many entities can be derived by different processes. For example, sequence composition can be derived from knowledge of a protein sequence, or by biochemical analysis of the purified protein.

Some nodes in Figure 2 (Sequence Profiles and Secondary Structure) are used to represent structured sets of entities (single, multiple, aligned, and consensus) reflecting whether the analysis has produced single, multiple, or aligned sequences or the consensus of an alignment. The node labeled Results of Mutagenesis Experiments represents biochemical and biophysical information that is potentially relevant to protein structure prediction (e.g., tritium trapping experiments), not represented by any other node. Finally, under Sequence Profiles are grouped all the 222[28] different protein sequence profiles. Figure 2 is presented to reflect the breadth of information that may be accommodated by the network architecture, but is not intended to be a complete empirical description.

Finally, although some of the knowledge sources identified in Figure 2 have been implemented as software or are trivial to implement, many are major areas of research in their own right (e.g., crystallography, interpreting CD spectra, and using 2D NMR distance constraints to predict tertiary structure). Our aim in using the abstract representation scheme is not to trivialize these areas of research, but to provide a conceptualization of how they may be formally integrated with other knowledge applicable to protein structure prediction and sequence analysis.

### Relations

There are two kinds of relations (links) between the entities shown in Figures 2 and 3: *constraints* (thinner lines), which are consistency requirements between entities, and *minimal preconditions* (thicker lines), which are associated with processes that relate entities.

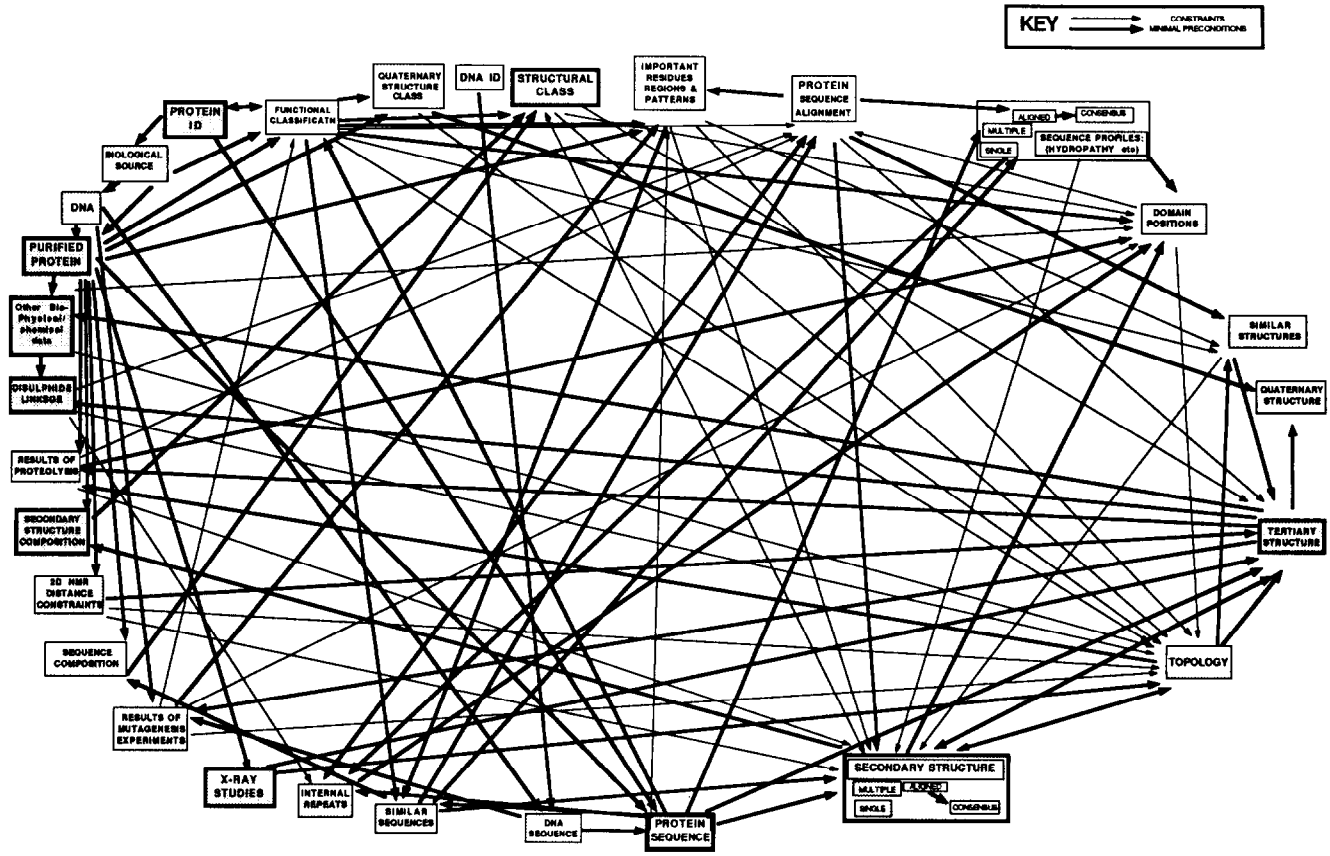


*Figure 2. Network diagram showing minimal preconditions and constraints among various entities employed in protein structure prediction*

***Minimal preconditions*** *A* is a minimal precondition for *B* if, under some circumstances, there is a process that can be used to derive an hypothesis for *B* from *A*. For example, similar sequences are a minimal precondition for an alignment. This is a minimal precondition because other information (additional preconditions) may need to hold simultaneously for the process that relates the associated entities to be applicable. For example, to derive a protein sequence from a eukaryotic DNA sequence the positions of exon–intron junctions must be known. Less trivially, predicting structural class from sequence composition or secondary structure composition assumes that the protein of interest has only one domain or that all domains have the same structural class (unlike papain, for example). Minimal preconditions and additional preconditions are individually necessary conditions for the execution of processes, and therefore the conjunction of the set of a minimal and additional preconditions are a sufficient condition. Additional preconditions are not shown in Figures 2 and 3.

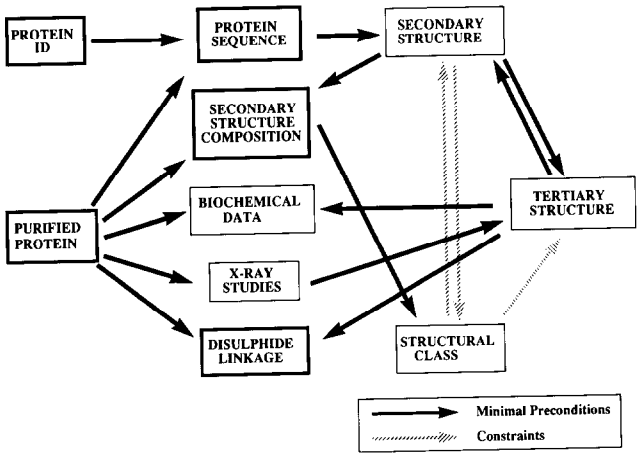


*Figure 3. Subnetwork of Figure 2 relevant to the prediction of HGH*

In general, if *A* is a minimal precondition for *B*, then *B* also constrains *A*. For example, since DNA sequence is a minimal precondition for protein sequence, knowledge of a protein sequence constrains the possible DNA sequences that could code for that protein.

Finally, when the minimal and additional preconditions for the process connecting two entities are all true, it is possible to execute that process. However, there is no guarantee that execution will generate a value for the target entity from the source entity. This is because success (e.g., finding >15% similarity in a similarity scan) depends on contextual factors (such as whether there is a sequence with this degree of similarity in the database).

***Constraints*** In contrast to minimal preconditions, which relate to processes connecting entities, constraints are consistency requirements between entities. *A* constrains *B* if the information contained in *A* limits the range of possible val-

**Table 2. Entities represented in Figure 2[a]**

| Entity | Description |
|---|---|
| Biological substance | |
| Biological source | Substance |
| DNA | Substance |
| Protein | Substance |
| Structural description | |
| DNA sequence | seq BASES |
| Protein sequence | seq AMINO ACIDS |
| Secondary structure | seq SECONDARY STRUCTURE ASSIGNMENTS |
| Topology | Set {qualitative sequence, orientation and adjacency relations} (super secondary structure is a subset of topology) |
| Tertiary structure | *xyz* coordinates |
| Quaternary structure | *xyz* coordinates of quaternary complex |
| Classifications and identifiers | |
| DNA ID | Name or database ID |
| Protein ID | Name or database ID |
| Functional classification | Set {substrate/cofactors/prosthetic group/E. C. no. etc.} |
| Tertiary structural class | Set {all-$\alpha$, all-$\beta$, $\alpha + \beta$, $\alpha/\beta$, irregular} |
| Quaternary structural class | Set {monomer, dimer, trimer, teteramer, etc.} |
| Results of biophysical and biochemical assays | |
| 2D NMR distance constraints | Set {pairs of residue distances} |
| Secondary structure composition | % $\alpha$ structure and % $\beta$ structure (parallel/antiparallel) |
| Results of proteolysis | Set {sequence positions of fragments} |
| Disulphide linkage | Set {linked cysteine pairs} |
| Domain positions | N and C terminal sequence positions |
| Results of x-ray studies | Electron density maps of varying resolution |
| Sequence composition | Absolute/relative composition of amino acids |
| Results of mutagenesis experiments | Interpretation of site directed, cassette, deletion and substitution mutagenesis |
| Results of other experiments | Various, see text |
| Results of database queries, sequence analyses and other software | |
| Internal repeats | Set {seq AMINO ACID/GAP} |
| Similar sequences | Set {protein IDs} |
| Alignment | Set {seq AMINO ACID/GAP} |
| Important patterns/regions | Residue identifiers or templates |
| Sequence profiles | seq VALUES |
| Similar structures | Protein IDS and domain/region identifiers |

[a] seq, sequence datatype.

ues or conformations of *B* in some situation, and *A* is not a minimal or additional precondition for *B*. Thus, disulfide linkage constrains an alignment because corresponding cysteine residues in the disulfide linkage should be aligned, and disulfide linkage is not a precondition for alignment. Similarly, the results of proteolysis can be used to constrain an alignment since cleaved regions will usually be in exposed loops which are unlikely to be conserved unless at an active site.

In general, the nature of constraints as requirements for

consistency means that the relation is symmetric; i.e., if $A$ constrains $B$, then $B$ also constrains $A$. So just as cleaved regions are not usually conserved in an alignment (unless at an active site), conserved regions in an alignment would not be expected to be cleaved. For simplicity, Figure 2 shows some constraints unidirectionally, where the direction shown indicates a move from more to less reliable information.

Complex interdependencies between entities can be modeled using the minimal precondition and constraint relations. For example, secondary structure prediction is a minimal precondition for secondary structure composition, which is, in turn, a minimal precondition for structural class. But secondary structure prediction is not a minimal precondition for structural class because its action in predicting structural class is predominantly through the assessment of secondary structure composition (Figure 3). It is, however, a constraint on structural class because, under some conditions (~15% $\alpha$ and ~15% $\beta$ structure), the particular secondary structure prediction can help disambiguate whether a protein domain belongs to the $\alpha/\beta$ or $\alpha + \beta$ structural classes by observation of whether the predicted $\alpha$ helices and $\beta$ strands alternate (as in TIM) or are clustered into groups (as in lysozyme).

## PROTOTYPE SYSTEM

A prototype system has been developed that illustrates how the network model can be employed in a knowledge-based support system for orchestration of sequence analysis and data acquisition for protein structure prediction. The system is written in PROPS2, a Prolog production system interpreter and knowledge programming language,[29,18] and runs on a SUN 3. The interface is written in C using SunView. This interface includes dynamic, walk-through menus and an interactive (clickable) facade. Executing (clicking) a facade item or selecting a menu item is equivalent to typing that item via the keyboard. For the purposes of demonstration it is assumed that the field of interest is confined to Figure 3. Other assumptions are described below.

The system has 6 modules (Figure 4). The *I/O module* accepts keyboard and mouse-based input and handles screen formatting. The *network description* (Figure 5) is a set of

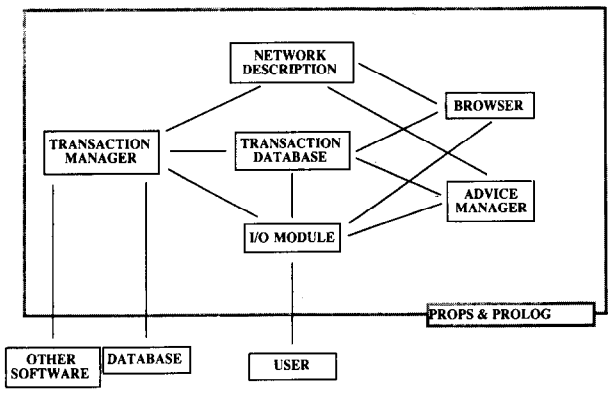


*Figure 4. Architecture of the prototype system. The network description is shown in Figure 5, while the operation of the other modules is demonstrated in Figures 6–10*

```
protein id is a minimal precondition of protein sequence by protein sequence database scan.
purified protein is a minimal precondition of protein sequence by protein sequencing techniques.
purified protein is a minimal precondition of biochemical data by biochemical assays.
purified protein is a minimal precondition of xray studies by crystallographic methods.
purified protein is a minimal precondition of secondary structure composition by circular
    dichroism spectra analysis.
purified protein is a minimal precondition of disulphide linkage by chemical methods.
protein sequence is a minimal precondition of secondary structure by secondary structure
    prediction techniques.
secondary structure is a minimal precondition of secondary structure composition by secondary
    structure frequency analysis.
secondary structure is a minimal precondition of tertiary structure by the Cohen packing
    algorithm.
secondary structure composition is a minimal precondition of structural class by secondary
    structure composition analysis.
tertiary structure is a minimal precondition of biochemical data by biochemical hypothesis
    generation.
tertiary structure is a minimal precondition of disulphide linkage by disulphide definition.
tertiary structure is a minimal precondition of secondary structure by kabsch/sander definitions.
xray studies is a minimal precondition of tertiary structure by electron density map solution
    techniques.

secondary structure is a constraint of structural class.
structural class is a constraint of secondary structure.
structural class is a constraint of tertiary structure.
```

*Figure 5. PROPS2 facts showing network description of Figure 3. This version of the prototype assumes only one method per process*

propositions that describe the relevant network (in this case Figure 3). The *transaction manager* maintains a *transaction database* that records sources of information and consistency between entities and uses this with the network description to determine which transactions are permissible at each stage of an interaction. The *browser* allows the user to browse features of the network description and the transaction database while the *advice manager* responds to other user queries illustrated below.

There are five formal transactions known to the transaction manager. These are *knowledge entry* (entering knowledge about the protein(s) of interest), *node derivation* (deriving values for one entity from another via the connecting process, e.g., deriving a secondary structure prediction from a protein sequence), *consistency checking* (determining the mutual permissibility of values for two entities), *node updating* (changing the value of an entity to accommodate new constraints or make it consistent with some other node), and *retraction* (retracting either user supplied information or withdrawing system processes).

Interactions have no fixed structure but might begin with the user entering information about a protein. This may be anything from simply the name or some other (database) identifier (e.g., human growth hormone) to a wide variety of biochemical/biophysical and interpreted data. In this demonstration, five pieces of information are entered corresponding to stage 1 in Table 1. These are the *protein id*, the *protein sequence*, the *disulfide linkage* (previously determined by biochemical means), the *secondary structure composition* (previously determined from CD spectra), and *purified protein* (Figure 6). Entering *purified protein* indicates that the user has a quantity of purified protein. On the basis of this information and the network description (Figure 5) the transaction manager uses rules such as that shown in Figure 7 to determine which transactions are possible. The set of data input by the user is initially assumed to be consistent by the system.

Figure 7 states that if some piece of information ($X$) is known and this information is a minimal precondition for the derivation of some other information ($Y$), which is not

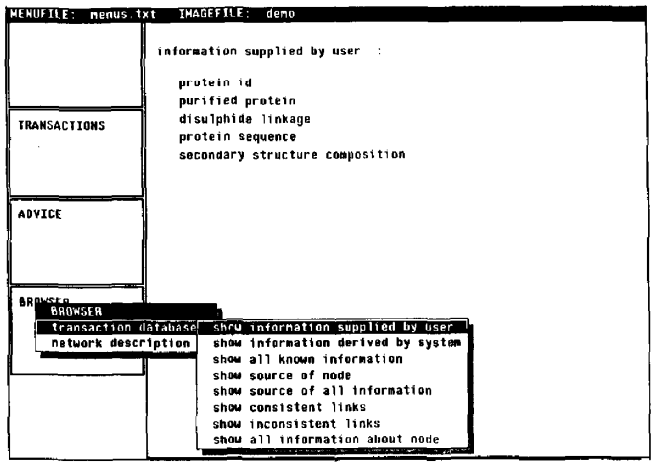


*Figure 6. Screen from interactive PROPS2 facade showing information supplied by the user at start of a session. The left side of the facade has three walk-through menus with functions of the browser, transaction manager, and advice manager modules of the prototype system (Figure 4). Selecting a menu item or clicking on an item displayed on the facade is equivalent to typing that item*

known, then it is possible to derive the unknown information ($Y$) from the known information ($X$) by the connecting process ($Z$). The term *derive* is used throughout this description for consistency. However, the process of deriving may variously be thought of as generating, creating, executing or predicting, as in predicting the secondary structure of a protein from its sequence. For this demonstration all additional preconditions are assumed to be true.

For the rule in Figure 7 to be executed, the three premises (the IF part) of the rule in Figure 7 must all be true. PROPS2 does not try to prove that *Y is not known* is true but employs *negation as failure*, a logic programming concept[30] in which all negated propositions (such as *Y is* not *known*) are assumed to be true unless explicitly contradicted by an affirmative proposition in the database. The rule in Figure 7 is executed, therefore, each time a match is found for an $X$ in *X is known*, and in *X is a minimal precondition of Y by Z* and where it is not true that *Y is known*. Therefore, if *protein sequence is known* is entered, the transaction manager uses the network description fact (Figure 5) *protein sequence is a minimal precondition of secondary structure prediction by secondary structure prediction techniques* to infer that *it is possible to derive secondary structure prediction from protein sequence by secondary structure prediction techniques*. However, should a secondary structure prediction subsequently be determined and the proposition *secondary structure prediction is known* become true, the third premise (Figure 7) will become false and *it is possible to derive secondary structure prediction from protein sequence by secondary structure prediction techniques* would be retracted. There are further rules in the system that determine the applicability of the other types of transactions.

Because of the separation of descriptive knowledge and the specification of how that knowledge is to be used, the system can operate and switch between data-driven mode of operation (data are supplied and possible actions requested) and goal-directed mode (goal is supplied and actions to achieve goal are requested), illustrated below.

```
if X is known
  and X is a minimal precondition of Y by Z
  and Y is not known
then it is possible to derive Y from X by Z
```

*Figure 7. A transaction manager rule for inferring possible derivations (uppercase letters are variables)*

## Data-driven interaction

Having entered what is known about the protein, the system is now asked to show the set of *possible transactions* (Figure 8*a*). These are the transactions that are possible given the network description, the current transaction database, and the definitions of the transactions. Four possible transactions are identified by the transaction manager: Three are simple derivations; the fourth (the derivation of a secondary structure prediction) has the constraint of consistency with the secondary structure composition. Clearly, the composition of secondary structure derived from CD spectra should be consistent with the proportion suggested by the secondary structure prediction.

Following the stages reported by Cohen and Kuntz[8] (Table 1) the derivation of structural class from secondary structure composition is simulated by executing *derive structural class from secondary structure composition*. The secondary structure composition of HGH was estimated to be 45–50% $\alpha$ structure and 0% $\beta$ structure; therefore, the structural class assigned was all-$\alpha$. As a result of this transaction, the set of possible transactions is changed to those shown in Figure 8*b*. The differences between Figures 8*a* and 8*b* are that derivation of structural class has been eliminated from the set of possible transactions and that structural class has been added as a constraint to secondary structure prediction.

Similarly, the prediction of secondary structure from protein sequence is simulated, taking into account constraints imposed by the secondary structure composition and the predicted structural class. Note that the all-$\alpha$ structural class assignment does allow small amounts of $\beta$ strand to be predicted. In this example, it is assumed that the constraints from the secondary structure composition and structural class can both be accommodated within the secondary structure predicted. In general, it is assumed that all known information that constrains an entity is applied when a value for that entity is initially derived. If this were not possible, the inconsistency between the constraint nodes (secondary structure composition and structural class), the generating node (protein sequence), and the associated process (secondary structure prediction techniques) would be recorded as an impasse and indicated to the user (see Figure 8*i*).

The new set of possible transactions is shown in Figure 8*c*. When predicting tertiary structure from secondary structure the system observes that both the disulfide linkage of


*Figure 8. Figures 8a–j show the textual output from the prototype system as a result of various user actions. For explanation purposes, the usual output (see Figure 6) has been simplified. Lines beginning with the bullet symbol are executable transactions. Boldface italic items are items that have not previously appeared on the screen. Items that are not boldface have already appeared on the screen. The prototype makes a number of assumptions for demonstration purposes, which are described in the text. These include the assumption that the field of protein structure prediction is limited to the nodes and relations shown in Figure 3. Figure 8a is a request for the set of possible transactions given the information shown in Figure 6. Figure 8b shows the new set following the prediction of structural class from secondary structure composition. Figure 8c shows the set of possible transactions following prediction of secondary structure. Figure 8d likewise shows the set of possible transactions following the prediction of tertiary structure from secondary structure using the Cohen helix packing algorithm. Figure 8e shows the consistent nodes at this stage in the interaction, while Figure 8f gives advice on how the accuracy of the predicted tertiary structure may be assessed. Figure 8g shows the set of possible transactions following the derivation of x-ray data by crystallography. At this point the consistency of the proposed tertiary structure is checked against the x-ray data and found to be inconsistent. Figure 8h shows the new set of possible transactions, while Figure 8i shows the system message in response to a request to the unexecutable request to update (make consistent) the predicted tertiary structure with the x-ray data. Figure 8j shows the possible retraction operations to remove this inconsistency*

the protein and structural class specific folding constraints should be taken into account. One such constraint imposed by Cohen and Kuntz, but subsequently demonstrated to be incorrect, was the assumption that all four helical bundles are right handed (HGH was shown to contain a left-handed four helical bundle).

Tertiary structure is now derived from the secondary structure prediction by Cohen's combinatorial algorithm, incorporating *constraints from disulfide linkage* and *structural class* to simulate the processes described in Table 1, steps 4–7. Cohen and Kuntz describe the incorporation of constraints as a sequential process of first generating a complete set of topologies from their secondary structure prediction and then reducing this set to exclude those topologies not consistent with the proposed disulfide linkage and structural class folding rule. In the demonstration the process is presented as a parallel operation. The resulting set of possible transactions is shown in Figure 8*d*.

Figures 8*e* and *f* show aspects of the functionality of the browser and the advice manager. Figure 8*e* is the browsers' response to a user query to show consistent links (which pairs of entities are consistent) in the transaction database. Figure 8*f* shows the advice manager's response to a query to give advice on how to assess the accuracy of the predicted tertiary structure. The advice is given in terms of transactions known to the transaction manager and is generated dynamically by the advice manager from planning rules, the network description, and the transaction database. Using general (sometimes recursive) planning rules, it is possible to construct complex advice. In reply to the specific query, the advice manager suggests two methods. First, x-ray data (an electron density map) can be produced by crystallography and its consistency checked against the predicted tertiary structure (note that this need not be a high-resolution electron density map). Second, the tertiary structure could be used to predict the outcome of various biochemical assays (e.g., proteolysis) and these predictions could then be assessed experimentally. For more complex networks such advice can assist scientists in the determination of methods for evaluating hypotheses about protein structure and function.

Following Table 1, the generation of x-ray data is now simulated and a new set of possible transactions is requested (Figure 8*g*). This set includes consistency checking, which gives the result as true or false. If predicted events were characterized by probabilities, consistency checking would involve the assignment of second-order probabilities. In this demonstration, it is assumed that the set of predicted topologies is not consistent with the x-ray data, as was the case for the prediction of HGH. This leads to a new option of *update predicted topologies using x-ray data* (Figure 8*h*). The idea is that by eliminating proposed structures that are inconsistent with the x-ray data, the remaining set will be consistent. Generally, resolving the inconsistency between two nodes may be possible by updating either. For example, if the existing functional classification of a set of similar sequences is not consistent with the individual known functions (e.g., all the sequences did not have the same function), it is possible to restore consistency by updating (changing) the set of sequences to eliminate those whose functions differ or update the functional classification of the set (e.g., make the classification more general). In the case of tertiary structure and x-ray data the transaction manager does not propose updating the x-ray data to account for the tertiary structure, because the network description is augmented by the fact *derivation of x-ray data is independent of tertiary structure* (not shown in Figure 5). This means that the tertiary structure cannot be used to change the x-ray data, though the tertiary structure can affect the interpretation of the x-ray data.

Attempting to update the set of predicted topologies using the x-ray data is unsuccessful (since all the proposed topologies are right-handed four helical bundles), yielding an impasse report (Figure 8*i*). Here it is clear that the set of information (both input and derived data about the particular protein and the knowledge bases that are used in the derivation) is inconsistent.

The advice manager is now asked for information to show how the inconsistency might be resolved. In general, resolving inconsistency involves retracting (changing or deleting) either the input information or the knowledge that is used to derive further results from the input information (Figure 8*j*). The report is generated from information in the transaction database and lists all the information that has a direct bearing on the inconsistency of the proposed topologies with the x-ray data. So, for example, there might have been a crystallographic error (such as the wrong space group) or the sample might be from some protein other than HGH. Alternatively, the Cohen helix packing algorithm or the disulfide linkage or the secondary structure prediction and so on may be wrong.

## Goal-directed interaction

Advice on checking the accuracy of the predicted tertiary structure requires some goal-directed reasoning (Figure 8*f*); however, the system can operate in a completely goal-driven manner. Given the initial data shown in Figure 6, Figure 9*a* shows the advice manager response to a user goal to predict tertiary structure. Two routes are shown: (1) predict secondary structure and then use the Cohen combinatorial algorithm and (2) by crystallography. Executing *derive secondary structure prediction by secondary structure prediction techniques* produces the augmented screen in Figure 9*b*. It is possible to switch between a data-driven and goal-driven modes and to access any system function at any time.

## DISCUSSION

The prototype system demonstrates some ways in which the entities and relations in Figure 2 can be utilized within a knowledge-based system to provide a computer-based support environment for protein sequence analysis and structure prediction. However, many other systems incorporating a network architecture can be envisaged. In addition to specification of further network elements and transactions, there are several routes to extension. The network model can be employed as part of an intelligent interface to molecular biology software for report construction, e.g., by running every applicable piece of analysis software and then pro-

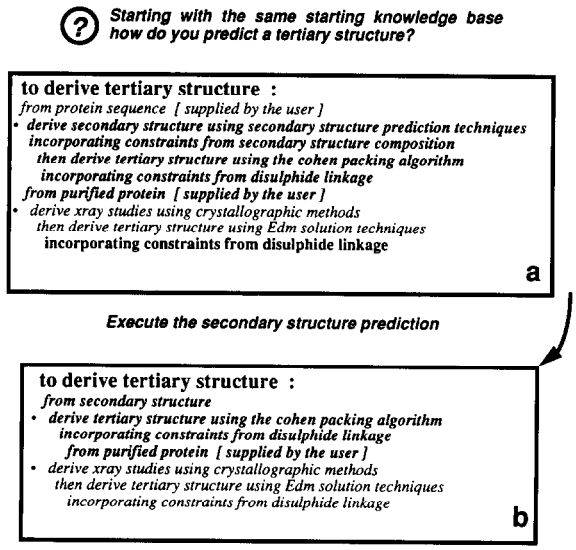


*Figure 9. Figures* 9a *and* b *demonstrate the goal-directed features of the prototype system. Figure* 9a *is the result of queries to show how tertiary structure may be predicted given the information in Figure 6, while Figure* 9b *gives the same information following the prediction of secondary structure*

ducing a report. Alternatively, higher-level advice could be implemented by identifying recommended, rather than just possible, transactions. This could be done in at least three ways: (1) by using a predetermined high-level strategy, such as that proposed by Taylor (Figure 1); (2) using dynamic utility assessment methods, assigning utility and reliability measures to entities and procedures and then maximizing some composite function,[31] or (3) by symbolic decision making.[32,33]

To illustrate symbolic decision making, it is clear from the demonstration that the order in which transactions are performed can affect the resulting state of the transaction database. For example, if *A* constrains *B* and *B* is derived before *A* is known, then *A* and *B* will not necessarily be consistent. However, if *A* is derived before *B* and then *B* is derived incorporating the constraints from *A*, then *A* and *B* will be consistent. It is therefore possible to write symbolic heuristics such as Figure 10, which utilize redundancy.

Currently an interface is being developed between the transaction manager and a subset of processes in Figure 2 which can be implemented as software. While some of the required software is publically available or has been written locally, there are two principal difficulties to be overcome in producing a complete operational system, relating both to software and biological knowledge.

Most existing software is inflexible to the imposition of constraints on the manner in which information is processed or solutions generated. Examples of programs that do permit incorporation of external constraints either through the use of user selected parameters or weights include the AMPS package,[34,35] which permits secondary structure and specific residues to be weighted in alignment, and the Robson secondary structure prediction algorithm,[36,37] where decision

```
if goal is X
  and A is a minimal preconditions of X by C
  and B is a additional preconditions of X by C
  and constraints of A include B
then derive B before A in context of goal X.
```

*Figure 10. Symbolic scheduling rule for constraining sequence of process execution*

constants can be used to incorporate information about structural class. To introduce constraints, some software must be reimplemented in a more flexible manner.

A more general difficulty is the nonstandardization of some of the knowledge and concepts of molecular biology (e.g., the definition of a structural domain) and the absence of a coherent functional classification of nonenzyme proteins. These are both unfortunate since functional information and knowledge of structural domains are potentially very powerful sources of information for reasoning about protein structure. This is because structural domains can be treated as independent folding units, while functional information is determined by and therefore constrains 3D structure. Any rapidly developing science is expected to revise and generalize the concepts it employs to characterize the entities of interest. However, the lack of standard definitions makes additional demands on software development. In particular, it increases the need for flexibility and modifiability in the way in which knowledge is represented. Fortunately, such flexibility is afforded by AI languages such as Prolog and PROPS. The network architecture presented here provides a strategy for tackling the general problem of integrating the diverse sources of knowledge that characterize many areas of science. For example, the network in Figure 2 could readily be extended to cover the management of the diverse sources of information used in the design of experiments in genetic manipulation and other areas of biology.[38]

The goal of this research was to develop an architecture for a knowledge-based system to assist scientists in the task of protein sequence analysis and structure prediction. It was argued that the required architecture should be capable of representing the large body of relevant information, flexible with respect to strategic knowledge, and modular to permit extension and refinement.

The network architecture presented in this paper satisfies this goal. In particular, it has been shown that the knowledge used in the prediction of HGH is amenable to representation in terms of the simple network architecture presented and elsewhere it has been demonstrated that it is possible to represent all the information employed in each of the case studies as a subgraph of the network in Figure 2.[39] Finally, it has been demonstrated that the network representation can be used as the basis for the development of a useful knowledge-based protein sequence analysis and structure prediction support system.

In terms of the broader goals of molecular biology, the question to be posed is whether the architecture can be used

to predict protein structures more accurately. In this respect, we argue that the concept of combining different types of information intelligently has already been empirically demonstrated by both the successful prediction of Crawford et al.[10] and the studies that have shown that the accuracy of secondary structure prediction techniques can be improved by the incorporation of additional constraining information such as the structural class of a protein,[36] using a family of aligned sequences,[40] top-down constraints from super secondary structural motifs,[41] and the judicious combination of different secondary structure prediction techniques.[10,42] The next task is to demonstrate that it is possible to go beyond pairwise coupling of information and improve the accuracy of protein structure predictions using many mutually constraining sources of information. A knowledge-based architecture is the most appropriate vehicle for such an endeavor.

## APPENDIX 1. SUMMARIZED PREDICTION OF ALPHA SUBUNIT OF TRYPTOPHAN SYNTHASE[10]

1. Initially, the **protein identifier, quaternary structure class ($\alpha_2\beta_2$ in all procaryotes, $(\alpha\beta)_2$ in ascomycetes)**, **functional class** (synthase), and **results of mutation studies** (two-site revertants and **biochemical modification studies** (single site and cross-linking) were known to Crawford *et al.*.
2. A set of 10 **similar sequences** were defined on the basis of functional class.
3. The similar sequences were **aligned** by a system of pairwise comparisons followed by adjustments by eye.
4. **Secondary structure predictions** were performed on the aligned sequences by two distinct methods.
   a. Garnier, Osguthorpe, and Robson (**GOR**) prediction performed on each aligned sequence. This prediction suggested that the protein belonged to the $\alpha/\beta$ class, so the DCs were modified and the prediction repeated. The individual predictions were maintained and a **consensus GOR** prediction was obtained by averaging the prediction profiles at each aligned position and taking the highest scoring state at that position.
   b. **Chou and Fasman** (CF) performed on each aligned sequence, ambiguities resolved by arbitrary rules. Consensus prediction obtained by taking best scoring state at each aligned position—just how is not clear . . . .
5. **Hydropathy** and **flexibility** profiles for each aligned sequence were determined and **average profiles taken over the alignment.**
6. **Consensus structure prediction** obtained by combining the GOR and CF predictions. Where ambiguity existed and CF differed from GOR then GOR was taken as correct since it is generally better at predicting $\alpha/\beta$ proteins. **Constraints** were applied to the prediction in the form of **Indels** being expected in loop regions and **hydrophobic, nonflexible** regions in $\beta$ strands.
7. **Hypothesis**: that the protein is an $\alpha/\beta$ TIM-like barrel proposed on the basis of the presence of 8 strands and alternating helices.
8. Proteins of known tertiary structure that fold like TIM were identified from the database/literature to check for consistency with the proposed model for the $\alpha$ subunit.
9. **Consistency checks**
   a. Overall length of the $\alpha$ subunit is consistent with observed TIM barrels (approximately 200 amino acids). *Note that Hurle et al.*[9] *quote $\alpha/\beta$ barrel as about 250 residues.*
   b. Lengths of secondary structure elements are similar to observed TIM barrels and consistent with other $\alpha/\beta$ proteins. Also, the *extra* helix is likely to be $\alpha 0$ since it is very short in the *Bacillus subtilis* $\alpha$ subunit.
   c. Scatter of secondary structure lengths about the mean is random. This is necessary to allow a symmetrical barrel to form.
   d. The most highly conserved region of the $\alpha$ subunit is predicted to be at the end of the molecule that has the C terminals of the $\beta$ strands. This is in agreement with the other known TIM-like barrel proteins. It also suggests that the contact site with the $\beta$ subunit is at this end of the protein, since it is known from biochemical experiments that the active site is composed of elements from *both* $\alpha$ and $\beta$ subunits. **(quaternary structure constraints).**
   e. Conserved residues exist at the same side of the molecule as the predicted active site. Some of these are polar and might be involved in the catalytic mechanism, while others are nonpolar and might be expected to form the contact site with the $\beta$ subunit.
   f. **Limited proteolysis** of the protein cleaves at sites that are predicted to be in loops or at the ends of secondary structures.
   g. Indolepropanol phosphate bound to the $\alpha$ subunit protects R179 from chemical modification, suggesting that this residue is close to the active site. This residue is also on the C-terminal face of the protein (tenuous).
   h. **Cross-linking** can be performed between two Cys residues (81 and 118) that in the model are on adjacent $\alpha$ helices and could be spatially close.
   i. Single-site and two-site **mutants** provide data consistent with the $\alpha/\beta$ barrel model.
   j. X-ray crystallography confirms that this structure is indeed a TIM-like barrel.
10. **Negative evidence:** The only evidence that does not support the view of a single $\alpha/\beta$ barrel are the two step unfolding studies. However, this is not particularly reliable information since other single-domain proteins (e.g., carbonic anhydrase) can be shown to have two partially folded intermediates.
11. **Other (difficult to classify) evidence** is that one site (R188) is accessible to trypsin even in the quaternary structure. This suggests that this residue is not buried in the interface.

## APPENDIX 2. SUMMARIZED PREDICTION OF ALPHA SUBUNIT OF TRYPTOPHAN SYNTHASE[9]

1. Initially known: **5 homologous protein sequences, 2-site revertant** mutagenesis **studies, chemical cross-linking and modification data, limited proteolysis and unfolding data, functional class** and **quaternary structure classification,** $\alpha_2\beta_2$.
2. The protein was purified and then **VUV CD** studies were conducted using reference spectra from proteins of known structural composition by the method of Hennesey and Johnson.[44] This study gave composition data of 41 ± 2% $\alpha$ structure, 12 ± 1% parallel $\beta$ structure, and 16 ± 3% $\beta$ turn.
3. The CD data were deemed most consistent with the $\alpha/\beta$ class of proteins. In particular, they were considered most consistent with an **$\alpha/\beta$ barrel!**
4. The five sequences were aligned (not explicitly stated). Overall identity across the five sequences was 25%.
5. **Secondary structure predictions** were performed for each sequence individually and a "majority rules" strategy was adopted (Cohen et al.[7] approach). The prediction procedure first assigned turns, then the interturn regions were assigned to $\alpha,\beta$ or unstructured by applying rules that include a knowledge of the *structural class*.

   Three consensus predictions were maintained at this stage, these had slight differences in the location and length of some of the secondary structures. Note that the rationale by which these particular predictions were chosen is not explicitly stated in the text.

   One prediction was made using the Chou and Fasman[43] procedure.

   One prediction was made on the basis of a *single domain* assumption. This prediction was subsequently eliminated, however, because the proteolysis and refolding experiments suggested that the protein had two independent domains. Not that this assumption discards the possibility of an $\alpha/\beta$ barrel structure.
6. **Sheet versus Barrel argument.** Proteolysis and refolding experiments suggest that there are two distinct folding units. If two domains *are* present then neither would be long enough to be a barrel. Hence the conclusion that this protein **is not a barrel,** but must be some sort of nonbarrel sheet.
7. **Fold predictions.** Secondary structure prediction produced four sets of predictions; all these were maintained at this level. Assuming only one sheet, all possible topologies were generated, then screened by applying several class specific constraints.
   a. Sequential $\beta$ strands have right-handed connections.
   b. Parallel $\beta$ sheets have no more than one change in winding direction.
   c. $\alpha$ helices evenly cover the sheet on both sides.
   d. Steric conflicts between $\alpha$ helices that connect strands are forbidden. These led to only *one* possible fold.
   e. The $\beta$ sheets must contain a hydrophobic stripe on both faces.
   f. The stripe is diagonal with respect to the strands when viewed flat.
   g. Hydrogen bonding is maximized in the sheet (within 2–3 of the maximum possible).

This led to a set of topologies. These were then further constrained by considering **strand alignment.**

Candidate structures were finally screened by considering **chemical cross-linking** and checking that the **hydrophobic patches** were conserved in *all* species.

8. For the single structure remaining, the $\alpha$ helices were packed against the sheet by first identifying suitable patches and the docking sites.
9. $\alpha$-Carbon coordinates were generated by reference to "ideal" structures.
10. **Validity** of the structure was considered by checking that the model was consistent with chemical cross-linking data, labeling, proteolysis, and protection studies. It appeared to be so; however, consistency checking with the x-ray structure showed that the model was wrong.

## ACKNOWLEDGMENTS

We are grateful to W. R. Taylor (NIMR) for making reference 12 available prior to publication and for comments on an early version of the prototype system. Thanks to John Fox (ICRF) and Mike Sternberg (ICRF) for comments on earlier drafts and Saki Hajnal for valuable technical support. We gratefully acknowledge support from I.C.R.F. (DAC,CJB) and the Science Engineering Research Council (DAC,CJR).

## REFERENCES

1. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 1987, **326**, 347–352
2. Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* 1987, **1**(5), 377–384
3. Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L. Knowledge based modelling of homologous proteins, part II. Rules for the conformation of substituted side-chains. *Protein Eng.* 1987, **1**(5), 385–392
4. Fox, J. A short account of Knowledge Engineering. *Knowledge Eng. Rev.* 1985, **1**(1), 4–14
5. Adeli, H., Ed. *The Handbook of Knowledge Engineering,* Vol. 2, McGraw-Hill, New York, 1989
6. Sternberg, M. J. E., and Cohen, F. E. Prediction of the secondary and tertiary structures of interferon from four homologous amino acid sequences. *Int. J. Biol. Macromol.* 1982, **4**, 137–144
7. Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, L. B., Ciardelli, T. L., and Smith, K. A. Structure-activity studies of interleukin-2. *Science* 1986, **234**, 349–352
8. Cohen, F. E., and Kuntz, I. D. Prediction of the three-dimensional structure of human growth hormone. *PROTEINS: Structure, Function and Genetics* 1987, **2**, 162–166

9. Hurle, M. R., Matthews, C. R., Cohen, F. E., Kuntz, I. D., Toumadje, A., and Johnson, W. C. Prediction of the tertiary structure of the alpha-subunit of tryptophan synthase. *PROTEINS: Structure, Function and Genetics* 1987, **2**, 210–224
10. Crawford, I. P., Niermann, T., and Kirschner, K. Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase. *PROTEINS: Structure, Function and Genetics* 1987, **2**, 118–129
11. Fishleigh, R. V., Robson, B., Garnier, J., and Finn, P. W. Studies on rationales for an expert system approach to the interpretation of protein sequence data. *FEBS Lett.* 1987, **214**(2), 219–225
12. Taylor, W. R., and Green, N. M. The predicted secondary structure of the nucleotide-binding sites of six cation-transporting ATPases leads to a probable tertiary fold. *Eur. J. Biochem.* 1989, **179**, 241–248
13. Rawlings, C. J., Taylor, W. R., Nyakairu, J., Fox, J., and Sternberg, M. J. E. Reasoning about protein topology using the logic programming language PROLOG. *J. Mol. Graphics* 1985, **3**(4), 151–157
14. Taylor, W. R. Protein structure prediction. In Bishop, M. J., and Rawlings, C. J., Eds. *Nucleic Acid and Protein Sequence Analysis, a Practical Approach*. IRL Press, Oxford, UK, 1987
15. Thornton, J. M. The shape of things to come. *Nature* 1988, **335**, 10–11
16. Rawlings, C. J. *Artificial Intelligence and Protein Structure Prediction*, Proceedings of Biotechnology Information '86. IRL Press, Oxford, UK, 1987, pp. 59–77
17. Frost, D., Fox, J., Duncan, T., and Preston, N. *Knowledge Engineering through Knowledge Programming: The PROPS2 Package*. ICRF Technical Report, 1986
18. Duncan, T. *The PROPS2 Reference Manual*. ICRF Biomedical Computing Unit Technical Report, 1986
19. Clark, D. A., and Barton, G. J. *Knowledge Engineering in the CARDS Project: A Review of Work in Progress*. ICRF, Biomedical Computing Unit Technical Report, 1988
20. Nii, H. P. Blackboard systems: The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine* 1986, **7**(2), 38–53
21. Nii, H. P. Blackboard systems: Blackboard application systems, blackboard systems from a Knowledge Engineering perspective. *AI Magazine* 1986, **7**(3), 82–106
22. Engelmore, R., and Morgan, T. *Blackboard Systems*. Addison-Wesley, Reading, MA, 1988
23. Erman, L. D., Hayes-Roth, F., Lesser, V. R., and Reddy, D. R. The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *Comput. Surveys* 1980, **12**(2), 213–253
24. Hayes-Roth, B., Buchanan, B., Lichtarge, O., Hewett, M., Altman, R., Brinkley, J., Cornelius, C., Duncan, B., and Jardetzky, O. PROTEAN: Deriving protein structure from constraints. In *Blackboard Systems*, Engelmore and Morgan, Eds. Addison-Wesley, Reading, MA, 1988
25. Terry, A. *The CRYSALIS Project: Hierarchical Control of Production Systems*. Technical Report HPP-83-19, Stanford University, Palo Alto, CA, 1983
26. Terry, A. Using explicit strategic knowledge to control expert systems. In *Blackboard Systems*, Engelmore and Morgan, Eds. Addison-Wesley, Reading, MA, 1988
27. Langlotz, C. P., and Shortliffe, E. H. Adapting a consultation system to critique user plans. In *Developments in Expert Systems*, M. J. Coombs, Ed. Academic Press, New York, 1984, pp. 77–94
28. Nakai, K., Kidera, A., and Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* 1988, **2**(2), 93–100
29. Fox, J., Frost, D., Duncan, T., and Preston, N. *The PROPS2 Primer*. ICRF Biomedical Computing Unit Technical Report, 1986
30. Clark, K. Negation as failure. In *Logic and Databases*, Gallaire and Minker, Eds. Plenum, New York, 1978
31. von Winterfeldt, D., and Edwards, W. *Decision Analysis and Behavioral Research* Cambridge University Press, Cambridge, UK, 1986
32. Fox, J. Symbolic decision procedures for knowledge based systems. In *The Handbook of Knowledge Engineering*, H. Adeli, Ed. McGraw-Hill, New York, 1989
33. Clark, D. A. Numerical and symbolic approaches to uncertainty management in AI: A review and discussion. *AI Rev.* 1990, **4**(2), 109–146
34. Barton, G. J., and Sternberg, M. J. E. A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol.* 1987, **198**, 327–337
35. Barton, G. J. Protein multiple sequence alignment and flexible pattern matching. In *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Methods in Enzymology*, Vol. 183, R. F. Doolittle, Ed. Academic, New York, 1989
36. Garnier, J., Osguthorpe, D. J., and Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 1978, **120**, 97–120
37. Gibrat, J.-F., Garnier, J., and Robson, B. Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* 1987, **198**, 425–443
38. Rawlings, C. J. Designing databases for molecular biology. *Nature* 1988, **334**, 477
39. Clark, D. A., Barton, G. J., and Rawlings, C. J. *Protein Structure Prediction: Knowledge Engineering in a Large Distributed Domain*. ICRF, Biomedical Computing Unit Technical Report, 1989
40. Zvelebil, M. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* 1987, **195**, 957–961
41. Taylor, W. R., and Thornton, J. M. Prediction of super-secondary structure in proteins. *Nature* 1983, **301**, 540–542
42. Biou, V., Gibrat, J.-F., Levin, J. M., Robson, B., and Garnier, J. Secondary structure prediction: Combination of three different methods. *Protein Eng.* 1988, **2**(3), 185–191
43. Chou, D. Y. and Fasman, G. D. Prediction of protein conformation. *Biochemistry*, 1974, **13**, 222–245
44. Hennessey, J. P., and Johnson, W. C. Information content in the circular dichroism of proteins. *Biochemistry*, 1981, **20**, 1085–1094