Proceedings AAAI Symposium on
"AI and Molecular Biology"
Stanford, Spring, 1990.

KNOWLEDGE-BASED ORCHESTRATION OF PROTEIN SEQUENCE ANALYSIS AND KNOWLEDGE ACQUISITION FOR PROTEIN STRUCTURE PREDICTION

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

Biomedical Computing Unit, Imperial Cancer Research Fund Laboratories,
London, WC2 3PX, UK.
*now at: Dept. of Biophysics, Oxford University.

ABSTRACT

The goal of this research is to produce more effective methods for protein sequence analysis and structure prediction through the use of knowledge-based techniques for orchestrating protein sequence and other analyses. We are developing a system (PAPAIN) that will provide intelligent assistance in manipulating and integrating diverse sources of information in a manner that will permit experimentation with hypothesis formation and reasoning styles. This paper describes foundational knowledge engineering studies, the resulting logical simulation and outlines current research.

1. INTRODUCTION

A complete understanding of protein function requires knowledge of its molecular structure. The direct investigation of protein structure by X-ray crystallography is, however, time-consuming and expensive. In consequence, scientists are increasingly turning to techniques for predicting the structure of a protein from its amino acid sequence. Techniques for protein structure prediction (henceforth PSP) fall into two classes: (i) theoretical methods (e.g. molecular dynamics and energy minimization) and (ii) empirical methods which combine sequence data with other information. From the perspective of the laboratory scientist, empirical methods are the most relevant and currently the most reliable is prediction using model building based on the close sequence homology with a protein whose structure has already been determined. When sequence identity is low, however, it is essential to seek other data to corroborate model-built structures. The range of proteins for which homologous structures can be found is limited, however, by the relatively small number of proteins whose structure has been adequately resolved. Furthermore, below about 15-25% sequence identity, model building is not applicable. Because of these restrictions to model-building, however, scientists are increasingly adopting *ad hoc* empirical methods that combine structural evidence from many different sources (e.g sequence and structure databases, experimental methods, protein sequence comparisons, protein sequence analysis) to arrive at a plausible structure prediction. Central to the success of these opportunistic methods is the knowledge and judgement of the participating scientist(s). Therefore it is not unusual for these methods to be unduly affected by biases such as failure to maintain consistency in the interpretation of data and failure to fully consider potentially conflicting information in the assessment of hypotheses.

The goal of our research is to develop more effective protein sequence analysis and structure prediction methods that acknowledge the central value of scientific expertise and diverse sources of structural evidence in generating and evaluating hypotheses about protein structure by the use of knowledge-based techniques. The potential value of knowledge-based systems for the scientific community has been widely discussed (e.g Hayes-Roth, 1987). Our technical goals are the development of knowledge-based tools which provide: (1) an interactive platform for the orchestration of protein sequence and other analyses; (2) intelligent assistance in the manipulation and integration of diverse sources of information; and (3) experimentation with hypothesis formation and lines of reasoning.

These goals encompass many issues that have received attention in AI research (e.g. integrating diverse knowledge sources, reasoning with uncertain evidence in a hierarchical problem space, spatial reasoning, scientific hypothesis formation and evaluation). This paper describes our foundation work on this problem, in particular the systematization of the types of knowledge used by scientists and how this knowledge can be exploited (Clark et al. 1990). We conclude with a brief

overview of work in progress as we move from a conceptual view to a more practical implementation designed to work with real data.

2. KNOWLEDGE ENGINEERING

The purpose of this study was to produce a conceptual description of knowledge (primarily descriptive) potentially relevant to PSP through the analysis of a set of publications that predicted protein structure without explicit use of model building techniques, augmented by the personal knowledge of GJB. The publications included the prediction of structures for interferon (Sternberg and Cohen, 1982), interleukin-2 (Cohen, 1986), human growth hormone (henceforth HGH, Cohen and Kuntz, 1987), α-subunit of tryptophan synthase (Hurle et al. 1987; Crawford et al. 1987), human epidermal growth factor receptor, (Fishleigh et al. 1987) and cation transporting ATPases (Taylor and Green, 1989). These papers typically employed sequence information in combination with biophysical and biochemical data, and used software analyses based on known 3D structures to arrive at a plausible tertiary structure.

Analysis involved identifying the logical organisation of each publication in terms of initial information, analyses/experiments reported, their interpretation, and arguments presented relating to cross validation and plausibility. The logical structure is typically the order in which journal papers are written. In general, the strategy adopted by authors was to try to produce the most consistent interpretation of the broadest range of data. Some of the papers, however, exhibited biases (such as failure to fully consider conflicting information) of the type documented in the literature on human judgement and decision making (Kahneman et al. 1982).

The analysis revealed many types of information that are relevant to PSP including biochemical and biophysical assays and the importance of topological reasoning, structural arguments based on functional similarity (Taylor and Green, 1989) and sequence motifs.

3. KNOWLEDGE REPRESENTATION

The diverse and rapidly changing nature of experimental techniques, analysis methods and data relevant to PSP, necessitate that an appropriate representational formalism for descriptive knowledge must be sufficiently modular and flexible to facilitate incremental extension and revision. To this end a network representation was developed (Clark et al. 1990). Figure 1 shows a subgraph of this network that pertains to the prediction of HGH. In this figure, nodes represent entities and links represent relations between entities which are either *minimal preconditions* (darker lines) associated with processes (the application of software, knowledge-based inference, biochemical and biophysical assays etc.) or constraints (thinner lines) which are requirements for consistency.

3.1. Entities

The entities in our network can be grouped into five categories (Clark et al. 1990). These are biological substance (e.g. the purified protein); protein structural descriptions at various levels of abstraction (e.g. amino acid sequence, secondary structure, protein topology, tertiary and quaternary structure); classifications and identifiers (e.g. functional, structural class and quaternary structural classification); results of biophysical and biochemical assays (e.g. proteolytic cleavage, chemical cross linking, site-directed mutagenesis, 2D-NMR, Circular Dichroism (henceforth CD) spectra and gel analysis); and derived data such as results of database queries, sequence analyses and the application of other software.

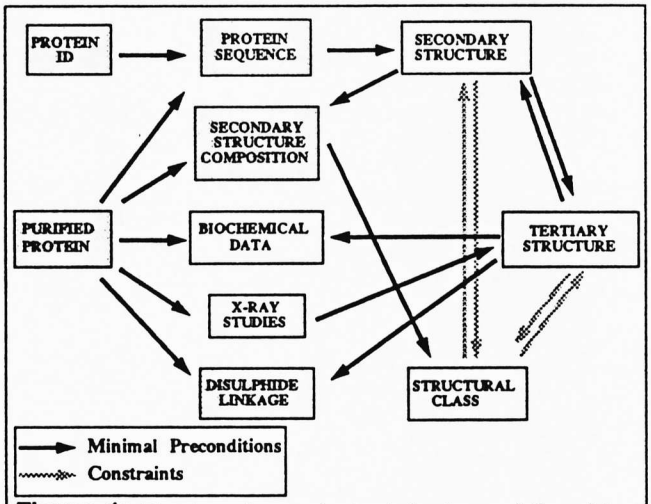

Figure 1: The dependency subnetwork for the prediction of human growth hormone (Cohen and Kuntz; 1987).

3.2. 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 a sequence alignment. Other (*additional*) preconditions (not shown in Figure 1) may need to hold simultaneously for the process that relates the associated entities to be applicable. The conjunction of the set of a minimal and additional preconditions are a sufficient condition for the execution of the associated process. It is also the case that 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, executing a process is no guarantee that it will generate an hypothesis for the target entity from the source entity. This is because the desired outcome (e.g. finding >25% global similarity in a sequence scan) will depend upon contextual factors.

3.3. Constraints

A constrains B if the information contained in A limits the range of possible values or conformations of B in some situation, and A is not a minimal or additional precondition for B. Thus, disulphide linkage constrains an alignment because corresponding cysteine residues in the disulphide linkage should be aligned, and disulphide linkage is not a precondition for alignment. Since constraints are requirements for consistency the relation is symmetric, (if A constrains B then B constrains A). So just as cleaved regions are not usually conserved in an alignment unless at an active site, so conserved regions in an alignment would not be expected to be cleaved, unless at an active site.

4. LOGICAL SIMULATION

To illustrate some of the ways in which the descriptive knowledge identified above could be employed in a knowledge-based support system for PSP, a program has been developed to simulate the logical steps in each of the prediction papers using the initial information and knowledge employed by the authors. The simulation system is user-driven but utilizes a constraint propagation mechanism for maintaining information about sources of information and consistency relations between entities. The simulation uses the system state to provide the user with advice concerning permissible operations and other types of advice. Figure 2 shows a dynamically generated system report showing all the information that has a direct bearing on an inconsistency that has been detected between the proposed structure of a protein and crystallographic data (based on the simulation of the prediction of HGH). Here the source of the inconsistency is traced to a set of data and analysis processes.

5. CURRENT ACTIVITIES

Our research is now focussed on a more practical system, PAPAIN (Protein Analysis and Prediction using Artificial INtelligence) which takes the use of the dependency network an important stage forward to that of providing knowledge-based assistance in the interpretation of protein sequence data.

cannot update tertiary structure with xray studies because of unresolvable inconsistency between:

(1) the tertiary structure
derived by the cohen packing algorithm
from secondary structure
[constrained by the structural class and disulphide linkage]
derived by secondary structure prediction techniques
[constrained by the structural class and secondary structure composition]
from protein sequence supplied by user

and

(2) the xray studies
derived by crystallographic methods
from purified protein supplied by user

Figure 2: Advice from the logical simulation indicating an inconsistency between the proposed tertiary structure and X-ray data.

In the PAPAIN system we are concentrating on the combined interpretation of data from secondary structure prediction methods, sequence profile analyses (e.g. charge and hydropathy), sequence pattern, topological and sequence motifs and secondary structure composition (as determined by CD spectra). This work is divided into two areas relating to (1) the development of a set of graphically oriented tools and (2) the development of formal techniques for constraint propagation, strategic reasoning and the management of uncertainty.

5.1. Graphical Tools.

Since the user has an important role in providing protein structural knowledge to the system, we are developing a variety of graphical user-interface tools that will enable the scientist to directly interact with the developing knowledge base.

We have previously shown how logic programming techniques can be used to represent the reasoning necessary to derive topological descriptions of protein structure (Rawlings et al, 1985), form the basis of a flexible query language (Rawlings, 1987) and be used by a graphical user interface for generating protein topology queries (Seifert and Rawlings, 1988). A set of graphical user interface tools have been developed to extend this work (Figure 3). These can be used to (1) initiate and orchestrate data analyses (*analysis tool*) (2) display and annotate sequence analysis profiles (*profile tool*) (3) display and interact with protein topology displays that control browsing and query formulation (*cartoon tool*)

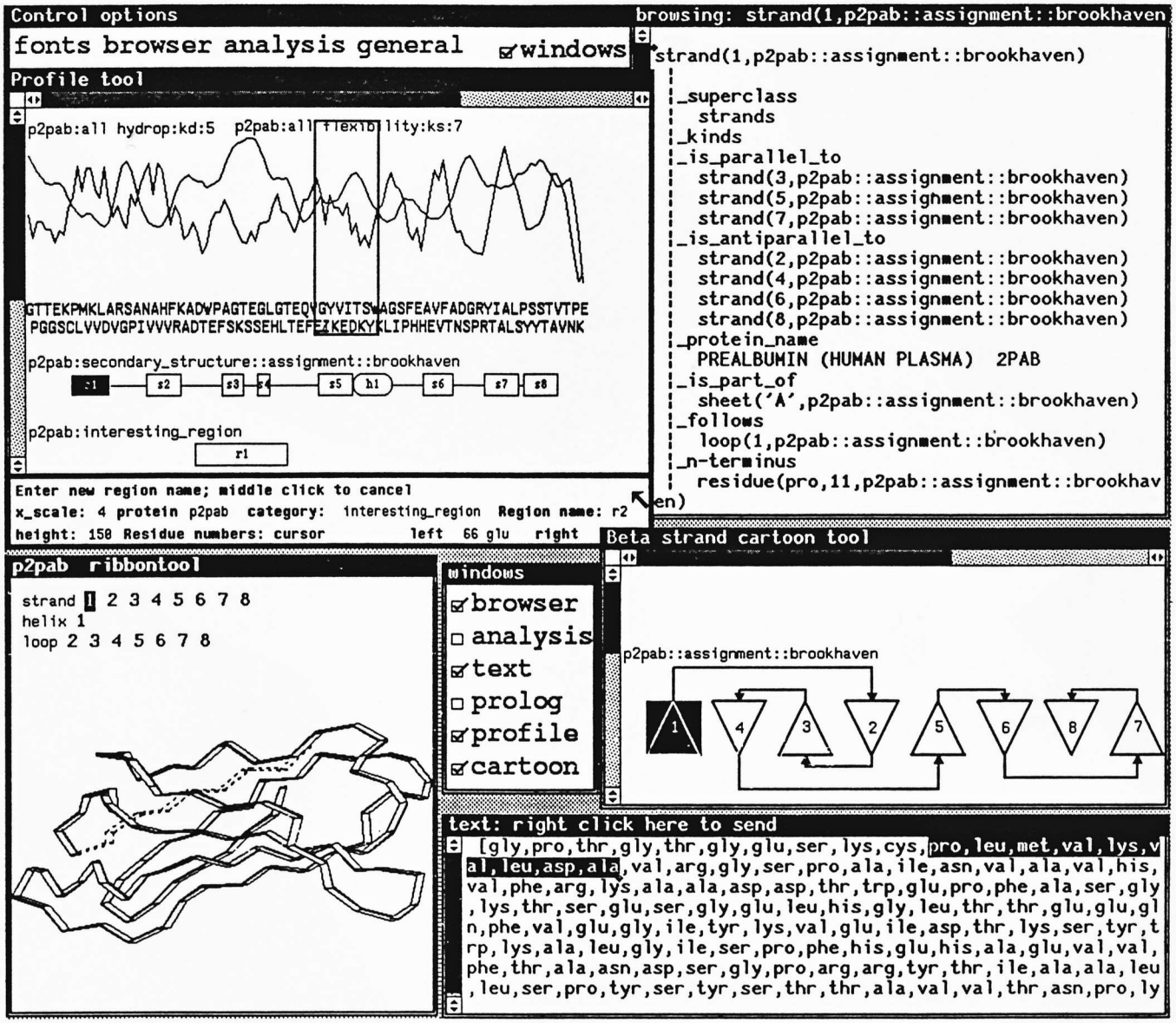

Figure 3: Graphical tools for knowledge-based orchestration of protein sequence analysis and knowledge acquisition.

(4) display protein secondary structures using a ribbon representation of the protein α-carbon backbone (*ribbon tool*) and (5) navigate frame-based generalisation hierarchies of protein structure and function (*browser*).

These tools are integrated in the sense (a) different windows provide complementary views of the same knowledge base (b) changes in the information displayed in one window causes displays in the other windows to be updated and (c) they can be employed in unison for data editing and knowledge acquisition. We are currently investigating interactive query construction using different windows to allow the user to construct complex constrained queries.

5.2. Constraints, Strategic Knowledge and Uncertainty

The importance of diverse sources of structural evidence in generating and evaluating hypotheses about protein structure has been demonstrated in a number of empirical studies that have shown that the accuracy of secondary structure prediction can be improved by the incorporation of constraints such as the structural class of a protein (Garnier, 1978), use of a family of aligned sequences, (Zvelebil et al. 1987), top-down constraints from super secondary structural motifs (Taylor and Thornton, 1983), and the judicious combination of different secondary structure prediction techniques (Biou et al. 1988). To extend this coupling of information in

a knowledge-based manner, we are currently investigating formal techniques for combining multiple sources of mutually constraining information.

The simulation described above provides advice on permissible uses of data and sequence analysis techniques for PSP and is based on a nonmonotonic model of uncertainty management. However, since our network represents only descriptive knowledge, it is strategically non-committal and it is possible to utilise this network representation (1) with different types of strategic and control knowledge and (2) different uncertainty management techniques. We are currently evaluating the use of predetermined high-level strategies (e.g. Taylor, 1987) and more general symbolic methods for managing uncertainty and providing strategic advice (Fox et al. 1990, Clark 1990).

References

Biou, V, Gibrat, J-F, Levin, J. M, Robson, B. and Garnier, J. (1988) *Secondary Structure Prediction: Combination of Three Different Methods.* Protein Engineering, 2(3), 185-191.

Clark, D. A. (1990) *Numerical and Symbolic approaches to Uncertainty management in AI: A Review and Discussion*, AI Review, 4(2), *to appear.*

Clark, D. A, Barton G. J. and Rawlings, C. J. (1990) *A Knowledge-Based Architecture for Protein Sequence Analysis and Structure Prediction.* J. Mol. Graphics, *to appear.*

Cohen, F. E, Kosen, P. A, Kuntz, I. D, Epstein, L. B, Ciardelli, T. L. and Smith, K. A. (1986) *Structure-Activity Studies of Interleukin-2.* Science 234, 349-352.

Cohen, F. E. and Kuntz, I. D. (1987) *Prediction of the Three-Dimensional Structure of Human Growth Hormone.* PROTEINS: Structure, Function and Genetics 2, 162-166.

Crawford, I. P, Niermann, T. and Kirschner, K. (1987) *Prediction of Secondary Structure by Evolutionary Comparison: Application to the alpha Subunit of Tryptophan Synthase.* PROTEINS: Structure, Function and Genetics 2, 118-129.

Fishleigh, R. V, Robson, B, Garnier, J. and Finn, P. W. (1987) *Studies on rationales for an expert system approach to the interpretation of protein sequence data.* FEBS Letters 214(2), 219-225.

Fox, J, Clark, D. A, Glowinski, A. and O'Neil, M. (1990) *Using predicate logic to integrate qualitative reasoning and classical decision theory.* IEEE Systems, Man and Cybernetics, *to appear.*

Garnier, J, Osguthorpe, D. J. and Robson, B. (1978) *Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins.* J. Mol. Biol., 120, 97-120.

Hayes-Roth, B. (1987) *Blackboard Systems*, In Encyclopedia of AI (Ed) Shapiro, S. C. John Wiley.

Hurle, M. R, Matthews, C. R, Cohen, F. E, Kuntz, I. D, Toumadje, A. and Johnson, W. C. (1987) *Prediction of the Tertiary Structure of the alpha-subunit of Tryptophan Synthase.* PROTEINS: Structure, Function and Genetics 2, 210-224.

Kahneman, D, Slovic, P. and Tversky, A. (1982) *Judgment under Uncertainty: Heuristics and Biases.* CUP.

Rawlings, C. J. (1987) *Artificial Intelligence and Protein Structure Prediction* Proceedings of Biotechnology Information '86 IRL Press, 59-77.

Rawlings, C. J, Taylor, W. R, Nyakairu, J, Fox, J. and Sternberg, M. J. E. (1985) *Reasoning about Protein Topology using the Logic programming language PROLOG.* J. Mol. Graphics 3(4) , 151-157.

Seifert, K. and Rawlings, C. J. (1988) *GRIPE - A Graphical Interface to a Knowledge Based System which Reasons about Protein Topology* In "People and Computers IV" (Eds) Jones, D. M. and Winder, R. British Informatics Society.

Sternberg, M. J. E. and Cohen, F. E. (1982) *Prediction of the secondary and tertiary structures of interferon from four homologous amino acid sequences.* Int. J. Biol. Macromol, 4, 137-144.

Taylor, W. R. (1987) *Protein Structure Prediction* In Bishop, M. J. and Rawlings, C. J. (Eds) "Nucleic acid and protein sequence analysis, a practical approach." IRL Press.

Taylor, W. R. and Green, N. M. (1989) *The predicted secondary structure of the nucleotide-binding sites of six cation-transporting ATPases leads to a probable tertiary fold.* Eur. J. Biochem 179, 241-248.

Taylor, W. R. and Thornton, J. M. (1983) *Prediction of super-secondary structure in proteins.* Nature 301, 540-542.

Zvelebil, M. J, Barton, G. J, Taylor, W. R. and Sternberg, M. J. E. (1987) *Prediction of Protein Secondary Structure and Active sites using the Alignment of Homologous Sequences.* J. Mol. Biol. 195, 957-961.