MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo 902 May, 1987

ARIADNE: Pattern-Directed Inference and Hierarchical Abstraction in Protein Structure Recognition

Richard H. Lathrop*, Teresa A. Webster**, Temple F. Smith**

* Artificial Intelligence Laboratory, Massachusetts Institute of Technology
** Molecular Biology Computer Research Resource, Dana Farber Cancer Institute, Harvard Medical School

ABSTRACT. There are many situations in which a very detailed low-level description encodes, through a hierarchical organization, a recognizable higher-order pattern. The macro-molecular structural conformations of proteins exhibit higher order regularities whose recognition is complicated by many factors. ARIADNE searches for similarities between structural descriptors and hypothesized protein structure at levels more abstract than the primary sequence, based on differential similarity to rule antecedents and the controlled use of tentative higher-order structural hypotheses. Inference is grounded solely in knowledge derivable from the primary sequence, and exploits secondary structure predictions. A novel proposed alignment and functional domain identification of the aminoacyl-tRNA synthetases was found using this system.

Notes:
(1) This paper will appear in *Communications of the ACM.*
(2) Ariadne was the Cretan princess who gave Theseus a ball of thread, by which he found his way out of the Labyrinth after slaying the Minotaur.

This paper describes research performed jointly at MIT's Artificial Intelligence Laboratory, and at Harvard Medical School's Molecular Biology Computer Research Resource in the Dana Farber Cancer Institute. Support for the MIT Artificial Intelligence Laboratory's research is provided in part by ONR contract N00014-85-K-0124. Support for the Molecular Biology Computer Research Resource's research is provided in part by NIH grant number RR02275. Personal support for the first author was furnished by an IBM Graduate Fellowship, and during the early stages of this research by an NSF Graduate Fellowship.

© Association for Computing Machinery 1987
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

1. INTRODUCTION.

This paper reports on the development of a hierarchical pattern-directed inference system for the ill-structured problem area of protein structure analysis. The system (ARIADNE) identifies the optimal match between a given complex pattern descriptor and genetic (protein) sequences annotated with various inferred properties, by abstracting intermediate levels of structural organization. Inference is grounded solely in knowledge derivable from the primary sequence, and exploits such weakly inferred properties as secondary structure predictions and hydrophobicity. The proposed aminoacyl-tRNA synthetase alignment and functional domain identification shown below is new and was found using this system with an hypothesized descriptor.

There are many situations in which a detailed low-level description encodes, through a hierarchical organization, a recognizable higher-order pattern. For example, in the microworld of VLSI integrated circuits, transistors are organized into inverters, inverters into register cells, register cells into register banks, and so on up to microprocessor. Another example occurs in board games such as chess, where a low-level description of which pieces occupy what positions encodes high-level descriptions such as "queen-side attack", through intermediate levels such as "pawn supports knight". Or again, one sub-problem in vision research involves the organization of low-level features such as "red patch", "curved edge", or "corner" into identifiable objects, and the situation of these objects into scenes. The common theme to these and other examples is that a few primitive types of low-level features encode a complex higher-order pattern by forming *complex relationships* with other low-level features.

Recognition of a hierarchical organization from low-level detail proceeds most naturally by hierarchical construction of the intervening patterns. Each instance of a pattern, when recognized in a low-level description, becomes available as a feature element for higher-order patterns. In this way a justifiable pyramid of inferences, each of manageable complexity, may connect the low-level features to the more abstract.

Hierarchical organizations and patterns also permeate the natural world. The organization of the biopolymers, proteins, is an important example. Proteins consist of tens of thousands of atoms in an ordered spatial arrangement of high inherent complexity. Protein structure, the focus of this study, has a number of identifiable hierarchical levels: the primary sequence of amino acids; locally regular secondary structure foldings of the primary sequence; groupings of these into super-secondary structures; the larger functional domains;

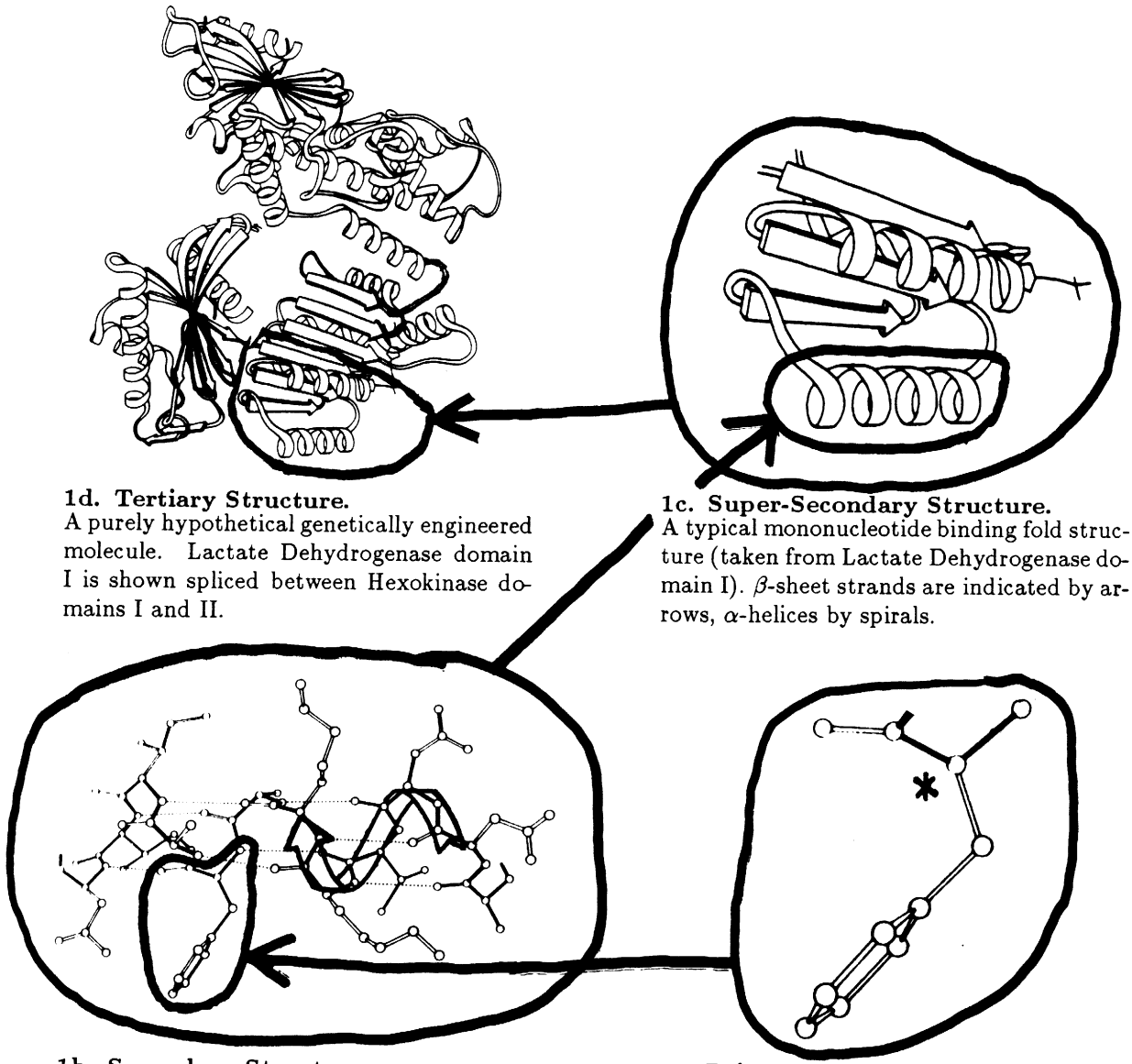
overall three-dimensional tertiary structure; and occasionally quaternary structure of multi-protein complexes (see figure 1 and the next section). Greatly complicating their analysis: protein three-dimensional structure is usually unknown; the processes by which amino acids form higher-order structures is poorly understood; pattern matching to known structures is inherently inexact due to mutations and various genetic rearrangements; patterns of interest are usually described only in terms of higher levels of organization; the applicable domain theory is incomplete, mostly heuristic, and incapable of directly predicting the desired higher-order groupings; and only weak and unreliable knowledge sources are available for feature generation at the lower levels of hierarchy construction. Kolata [22] terms the problem of inferring protein structure from protein primary sequence, "cracking the second half of the genetic code".

2. PROTEIN STRUCTURE.

"Genes are why we aren't cats." This simple truism expresses the fact that within the DNA sequence are encoded the instructions for building and regulating all biochemical hardware in living organisms. Proteins are one of the most important classes of encoded molecules. Each protein is a string written in a twenty-character alphabet of amino acid molecules. Enzymes are the proteins which control biochemical reactions, and thus indirectly most biological activity. Understanding biological activity requires an understanding of protein function, and this in turn is intimately linked to protein structure. A quite lucid exposition of basic protein structure is given by Richardson [36]. The general problem of inferring protein structure from primary sequence is summarized by Kolata [22]. The reader already broadly familiar with molecular biology may skip to the next section.

The protein string folds up in solution into a complicated globular three-dimensional shape, directly determined by the specific linear string of amino acids [2] (primary sequence, figure 1a). Regions of the primary sequence which fold into locally regular arrangements (α-helices, β-sheet strands, and β-turns) are termed secondary structures (figure 1b). Groupings of these often compose higher-order folding patterns known as super-secondary structures (figure 1c), which are less well-defined than the secondary structures. Enzymatically active sites (often cavities) may be composed of super-secondary structures, or may occur between larger protein sub-units known as domains. The full three-dimensional arrangement of the protein is termed tertiary structure (figure 1d). Occasionally multi-protein complexes assemble, forming quaternary structure.

The three-dimensional shape of a protein directly determines its biochemical activity.



1d. Tertiary Structure.
A purely hypothetical genetically engineered molecule. Lactate Dehydrogenase domain I is shown spliced between Hexokinase domains I and II.

1c. Super-Secondary Structure.
A typical mononucleotide binding fold structure (taken from Lactate Dehydrogenase domain I). β-sheet strands are indicated by arrows, α-helices by spirals.

1b. Secondary Structure.
A typical α-helix (residues 40-51 of the carp muscle calcium-binding protein). The helical coil passes along the backbone chain (darkened) with a periodicity of 3.6 to 3.7 residues.

1a. Primary Structure.
An amino acid, phenylalanine (three-letter code Phe, one-letter code F). Spheres represent atoms and rods represent chemical bonds. The alpha carbon is indicated by an '*'. The backbone is darkened.

Figure 1. Protein Organization.
The primary sequence is the linear chain of amino acids; it determines the helices, sheets and turns of secondary structure; the super-secondary groupings of these into biochemically active sites; the tertiary three-dimensional structure of the entire protein; and the quaternary structure of multi-protein complexes which sometimes form. These figures have been adapted (by permission) from a quite lucid presentation of protein structure by Richardson [36].

At enzymatically active sites the local surface structure conforms closely, like a glove to a hand or a lock to a key, to one or more of the chemicals involved; and a few key local amino acids influence the reaction. This enzymatic catalysis may result in a reaction occurring over a million times faster than in the absence of the protein. Location of active sites or cavities is important both for understanding the basic biochemistry of a protein and also for genetic engineering, which may be used to alter or combine sites to make a more effective pharmaceutical or a more useful industrial enzyme.

A protein's primary sequence can be easily discovered[1], in contrast to full tertiary structure determination from x-ray crystallography which is difficult and slow (if possible at all). Frequently the only structural information available about a potentially interesting protein is its amino acid sequence (figure 2a), and this will increasingly come to be the case in the future due to advances in sequencing technology. Although the primary amino acid sequence contains all the information necessary to specify the complete three-dimensional structure (figure 2b), the determinants of protein structure and function are unfortunately very poorly understood [22]. Quantum mechanics provides a solution in principle, but the computation is impractical for large proteins [26].

3. PREVIOUS PATTERN MATCHING.

In the absence of rigorous and tractable domain theory, prediction and exploration of protein structure are often approached by methods which compare primary sequences (reviewed by Waterman [41] and Sanoff [37]). Proteins with a substantial amount of primary sequence similarity invariably have similar functions and higher-order structures [10], with active sites found in corresponding regions. When part of a poorly-understood biological sequence is found to be similar in some respects to another better-understood one, an analogical[2] inference may map knowledge from the better understood case. These similarities have often had important and unexpected ramifications, as when human growth hormones were found to be similar to an oncogene (cancer related gene) [12].

Computer approaches to comparing biosequences have included finite-state grammars, regular expression matching, measures of "edit distance", exact string matches, and metric similarity measures [37, 41]. Most are designed to apply equally well to either the protein amino acid alphabet or the DNA nucleotide alphabet. These approaches have led to important advances, but have typically suffered from one or more of: failure to handle sequence

[1] Often indirectly, by determining the DNA sequence of the gene encoding the protein's amino acid sequence.

[2] Or homological. A homology is a similarity which arises from shared evolutionary history.

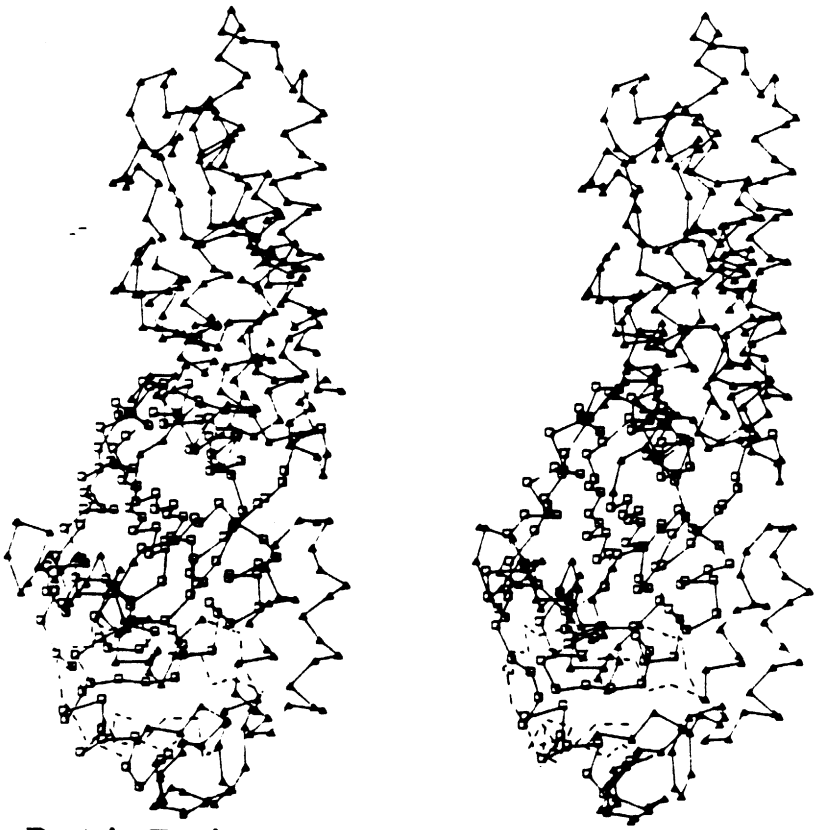


Figure 2b. Protein Tertiary Structure (Stereoscopic View).
The three-dimensional form of *E. coli* methionyl-tRNA synthetase after Zelwer et al. [50] (by permission). Tertiary structure is available for only one other synthetase (*B. stearothermophilus* tyrosyl-tRNA synthetase). Only the α carbon main chain atoms are shown. Square markers depict the "nucleotide binding domain". By focusing on the page while looking at infinity, it is possible to visually align the images stereoscopically. Alternatively, stereoscopic glasses may be employed.

T N V A K K I L V T C A L P Y A N G S I H L G H M L E H I N A D V W V R Y N
R M R G H E V N F I C A D D A H G T P I M L K A N N L G I T P E N M I G E M
S N E H N T D F A G F N I S Y D N Y H S T H S E E N R N L S E L I Y S R L K E N
G F I K N R T I S N L Y D P E K G M F L P D R F V K G T C P K C K S P D N Y
G D N C E V C G A T Y S P T E L I E P K S V V S G A T P V M R D S E H F F F D
L P S F S E M L N A W T R S G A L N E N V A N K M N E W F E S G L N N W D I
S R D A P Y F G F E I P N A P G K Y F Y V W L D A P I G Y M G S F K N L C D
K R G D S V S F D E Y W K K D S T A E L Y H F I G K D I V Y F H S L F W P A
M L E G S N F R K P S N L F V H G Y V T V N G A K M S K S R G T F I K A S T
W L N H F D A D S L R Y Y Y T A K L S S R I D D I D L N L E D F V N R V N A D
I V N K V V N L A S R N A G F I N K R F D G V L A S E L A D P N L Y K R F T D
A A E V I G E A W E S R E F G K A V R E I M A L A D L A N R Y V D E N A P W
V V A K N E G R D A D L N A I A N W G I N L F R V L M T Y L K P V L P K L T
E R A E A F L N T E L T W D G I N N P L L G H K V N P F K A L Y N R I D M R
N V E A L V E A S K E E V K A A A A P V T G P L A D D P N D G C G R H D R V
V D S G S K

Figure 2a. Protein Primary Structure (One-Letter Residue Abbreviations).
The 581 amino acid sequence for *E. coli* methionyl-tRNA synthetase [50]. This encodes the same information as figure 2b. ARIADNE's inferences are grounded solely in knowledge derivable from similar primary sequences.

element degeneracies; lack of a hierarchical organization in both pattern and biosequence representation; inability to perform a desired action easily upon noticing a match; an inflexible description language framework; and especially, difficulty in using hypothesized secondary structure predictions (or other weak, unreliable knowledge sources). The dynamic programming biosequence comparative methods [41] are currently used for finding similarities between various biosequences. Advances included the easy accommodation of partial similarities, inexact matches, and arbitrary length gaps. While the semantics of partial similarity used in these approaches are desirable, most of the problems mentioned above remain.

PLANS [1, 9] is a rule-based expert system successfully used to look for turns, and pioneered the use of a flexible recursive hierarchical pattern-matching language developed specifically for biosequences. PLANS was important because it showed the power and utility of a symbolic pattern descriptor. However, though the pattern definitions were hierarchical, the protein representation was not, making it difficult to exploit the secondary and super-secondary level information. Also, inference was based on exact matches to rule antecedents formed from regular expressions.

Gascuel and Danchin [18] successfully applied machine learning techniques to construct primary sequence descriptors which discriminate between prokaryotic (*E. coli*) and eukaryotic (human) signal sequences of exported proteins. They demonstrated the biological utility of procedurally-defined primitive descriptors, as well as induction of appropriate descriptors directly from data. (For a discussion of artificial intelligence and molecular biology see Friedland [16].)

Hayes-Roth et al. [20] are exploring a constraint-based approach to inferring the protein three-dimensional structure directly. This approach is not directly comparable with ours because the inferences are not derived solely from primary sequence information (the NMR used requires complex equipment and analysis), and because a specific active site is not identified (rather, many possible tertiary structures are returned). However, the initial results are interesting.

4. ARIADNE.

The major limitation of current biosequence comparative methods is that they require substantial primary sequence similarity in order to make inferences about protein structure. Although similar primary sequences generally indicates a similar folded conformation, the converse does not usually hold [25]. The problem occurs because secondary and super-

secondary structures are important in forming required spatial configurations, but do not often exhibit recognizable primary sequence patterns. PLANS [1, 9] and the method of Gascuel and Danchin [18] partially address this by allowing more complex patterns of primary sequence elements. ARIADNE facilitates direct expression and manipulation of higher-order structures, allowing direct use of secondary structure predictions and thus a search for similarities at a higher level than primary sequence.

A biologist first hypothesizes a protein structure of interest (see figure 3a) based on biochemical knowledge. This three-dimensional structure is "unfolded" to form a pattern descriptor as a sequence of primary and secondary structure elements (figure 3b). It is often convenient to describe this in terms of hierarchical groupings (figure 3c). ARIADNE receives as input the pattern descriptor and also the protein primary sequence (figure 4a) overlaid with predicted secondary structures (figure 4b). ARIADNE's biological structure knowledge is encoded in a number of pattern/action inference rules: an antecedent which describes a relationship between structural elements, and a consequent which hypothesizes the presence of a higher-order structure (see figure 5). The rules solely address structural organization, with as yet no "expert rule-of-thumb" knowledge of general biochemical heuristics. The target protein is searched for regions which are plausibly similar to the rule antecedent. When the rule fires its consequent typically creates a new entry in the overlay of predicted structures (compare figures 3c, 4c, and 5 for the Gly+helix). The new entry can enable the firing of subsequent rules, allowing a justified pyramid of manageable inferences to support the final hypothesized structure (figures 4c-e).

The power of pattern-directed inference (e.g., rule-based expert systems) is well known [11, 40], as is its applicability to molecular biology [15]. One of the first such systems ever constructed (DENDRAL [28]) also performed the task of chemical structure recognition. However, we allow flexible rule invocation based on a controllable degree of partial pattern similarity. This is implemented by an A* search [45] through the space of target protein subsequences. Much of our framework for abstraction manipulation comes from research into precedent-based inference [46, 47], clichés [35], and organization of active agents into hierarchies and other computational structures [31].

ARIADNE is implemented in LISP on a Symbolics 3600[3]. Design and construction of the basic research environment required roughly nine person-months of collaborative effort between a molecular biology domain expert and an artificial intelligence researcher.

[3] A trademark of Symbolics, Inc.

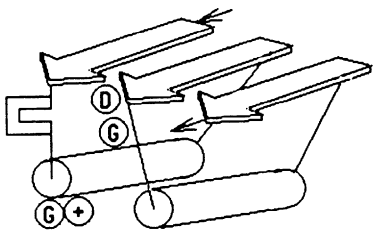


Figure 3a. Schematic of the Mononucleotide Binding Fold-like Structure.
β-sheet strands are represented by arrows, α-helices by cylinders, and β-turns by angular bends.

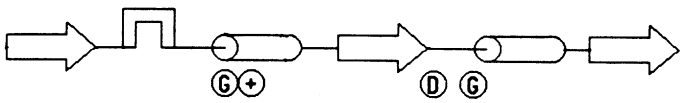


Figure 3b. The Mononucleotide Binding Fold Unfolded into a Linear Sequence.
The first β-sheet/β-turn/α-helix/β-sheet sequence will form the basis of the structural descriptor used in this paper. Key amino acids have been labeled.

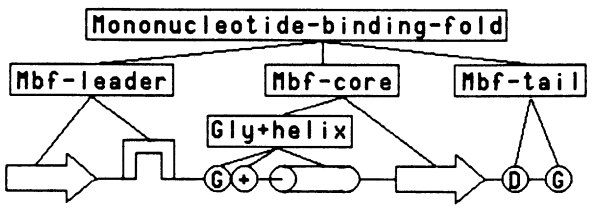


Figure 3c. The Unfolded Mononucleotide Binding Fold as Hierarchical Groupings.
It is often convenient to be able to describe a structure in terms of intermediate levels.

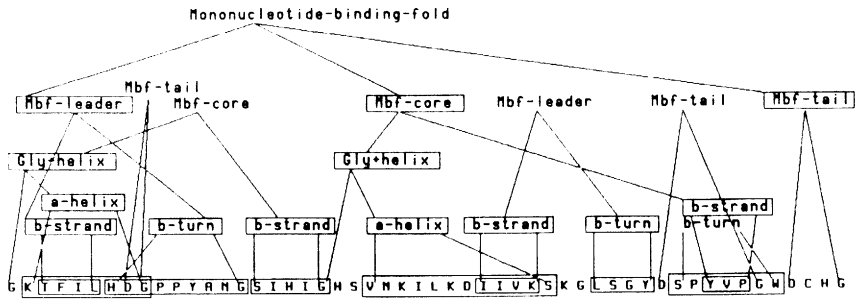


Figure 4e. E. coli Ile-RS (residues 48-99 of 939 residues)

(Final prediction constructed by ARIADNE.
No other instances of Mononucleotide-binding-fold are predicted.)

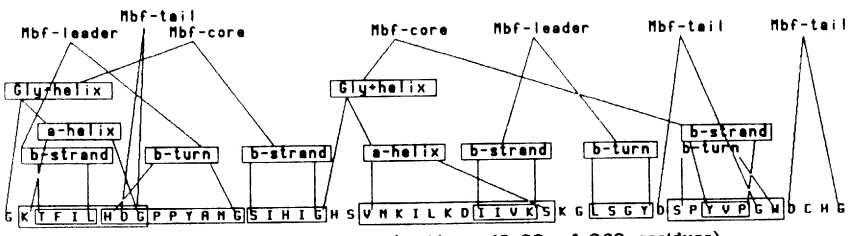


Figure 4d. E. coli Ile-RS (residues 48-99 of 939 residues)

(intermediate predictions constructed by ARIADNE)

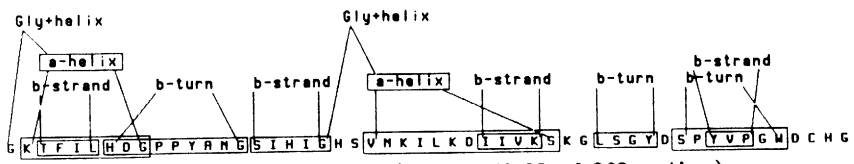


Figure 4c. E. coli Ile-RS (residues 48-99 of 939 residues)

(Intermediate predictions constructed by ARIADNE)

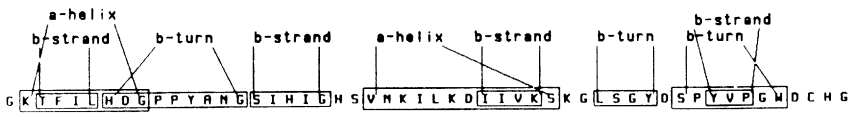


Figure 4b. E. coli Ile-RS (residues 48-99 of 939 residues)

(Chou & Fasman predictions [6, 35] input to ARIADNE)

G K T F I L H D G P P Y A M G S I H I G H S V M K I L K D I I V K S K G L S G Y D S P Y V P G W D C H G

Figure 4a. E. coli Ile-RS (residues 48-99 of 939 residues)

(Primary sequence input to ARIADNE)

Figure 4. Hierarchical Inference in ARIADNE.

```
(defpattern Gly+helix "Gly, +, helix"
  (pattern
    `(a-helix
       (near-front-of-prev
         :search-for
         ((G :score-if-mismatched ,-infinity)
          (or :amino-acids (C K H N Q R)
              :score-if-mismatched ,-infinity))
         :start-offset -5.
         :stop-offset +5.)))
  (action '((abstract-group))))

(defpattern MBF-LEADER "Introductory structures"
  (pattern '(b-strand
             (allow-overlaps :max-overlap 1.)
             (spacer :min 0 :max 4)
             b-turn))
  (action '((abstract-group))))

(defpattern MBF-CORE "Center Gly+helix, strand"
  (pattern `(Gly+helix
             (allow-overlaps :max-overlap 3.)
             (spacer :min 0 :max 11)
             b-strand))
  (action '((abstract-group)
            (record-in-buffer))))

(defpattern MBF-TAIL "Trailing key amino acids"
  (pattern `((D :score-if-missing -.5)
             (spacer :min 2 :max 2)
             (G :score-if-missing -.333)))
  (action '((abstract-group))))

(defpattern MONONUCLEOTIDE-BINDING-FOLD
            "Hypothesized MBF"
  (pattern '(MBF-LEADER
             (spacer :min 3 :max 3)
             MBF-CORE
             (spacer :min 0 :max 11)
             MBF-TAIL))
  (action '((abstract-group)
            (record-in-buffer)
            (expunge-overlaps)))
```

Figure 5. Composite Pattern Type Definition.
These would typically be written by the users of the system, creating patterns using primitives defined for them.

5. PREDICTING SECONDARY STRUCTURE.

Lacking the ability to perform a full quantum mechanical minimum energy analysis of all atoms as a function of their three-dimensional positions, the knowledge sources which connect the primary sequence to predictions of secondary structure α-helices, β-strands, and β-turns, are necessarily uncertain heuristics. Because the "best" indicators of secondary structure have surely not yet been developed, ARIADNE is designed to exploit a wide variety of potential sources:

1. Complex primary sequence patterns which represent secondary structure elements (for example, PLANS [1, 9]).

2. Output from any of several ancillary secondary structure prediction programs, discussed below.

3. Transforms of the primary sequence into a different representation, such that observable low-level features in the new representation are expected to be correlated with secondary structures (for example, hydropathy and hydropathy moment profiles [13, 12]).

4. Biochemical tests which indicate secondary structures experimentally (for example, NMR-based approaches [20]).

In actual practice we try to use two or more sources, to increase predictive accuracy.

For purposes of the discussion in this paper, however, the secondary structure α-helices, β-strands and β-turns were predicted solely by the ancillary program PRSTRC [34] based on the Chou and Fasman pseudoprobabilities [7]. There are several semi-empirical, heuristic methods for predicting secondary structure from primary sequence [7, 17, 27] and tertiary structure from secondary structure [8, 32]. The accuracy of most secondary structure prediction methods is only about 50-70% for α-helices and β-strands, and about 90% for β-turns, when compared to X-ray determined structures. The Chou and Fasman method of pseudoprobabilities appears to indicate a *local possibility* for secondary structure formation. Multiple overlapping predictions (which are mutually inconsistent) are often generated, but unfortunately without the ability to accurately choose between them. We set the PRSTRC parameters to optimally predict the actual secondary structures in the two synthetases with known tertiary structure. We also retained *all* multiple predictions. Thus if a given

secondary structure actually does exist in the protein, it is quite likely to be predicted; but, many spurious predictions are generated as well (see figures 4a-b).

6. PATTERN AND TARGET OBJECTS.

Input to ARIADNE consists of the primary sequence, any secondary structure predictions, and patterns describing the structure of interest. The primary sequence forms protein target objects, initially in a linear chain[4] (figure 4a). This is immediately overlaid with additional target objects (figure 4b) representing secondary structure predictions. Thereafter ARIADNE manipulates pairings consisting of a pattern p and a group of target objects $\{t_1, \ldots, t_n\}$. Each pair $m = (p, \{t_i\})$ represents an hypothesis that the group of target objects $\{t_i\}$ supports (or is similar to) the pattern p as parameterized. The pair m has an associated measure, $\sigma(m)$, of the similarity of p to $\{t_i\}$. Typically, a single new target object is created for each pair showing a positive similarity (figures 4c-e).

Viewed from top-to-bottom, the added target objects impose a hierarchical organization. Viewed from left-to-right they impose a lattice structure because of the partial ordering, "followed-by", inherited from the underlying linear chain. Pattern recognition consists of exploring alternate pathways through the lattice structure. For example, in figure 4b the target object representing the first lysine (the first "K" in "G K T F ...") may be followed either by a threonine object ("T") or by an object representing a β-strand prediction. The β-strand object, in turn, may be followed either by a histidine object ("H") or by a β-turn object. This permits structural elements (at any level) to be manipulated and searched as a unit, independent of their actual length or composition, in a way that is difficult or impossible in most existing biosequence analysis approaches.

7. PRIMITIVE AND COMPOSITE PATTERNS.

The pair $m = (p, \{t_i\})$ is treated differently depending on whether p is a composite or a primitive pattern. Composite patterns (e.g., branch nodes of figure 3c) are defined in terms of a pattern descriptor, which specifies a group of component objects and relationships. Primitive patterns are atomic (in the computational, if not the chemical, sense).

Primitive patterns usually appear only as components in higher-order pattern descriptors. They include the twenty primitive amino acids and various classifications (positively charged, hydrophobic, H-bond donors, etc.); several spacer, overlap, positioning, and containment operators; primitive graph features such as peak, valley and slope; and so forth.

[4]Observe that different micro-worlds would imply different organization, e.g. for VLSI structures the underlying relationship is topological connectivity rather than linear chain [24].

Their match behavior is governed by attached procedures which directly inspect and manipulate the target objects. Our overall goal is a declarative language of protein structure knowledge, but the ability to escape into procedural constructs facilitates exploration of which declarative forms may ultimately prove useful.

Composite patterns (see figure 5) possess a *pattern descriptor,* which is a declarative representation of the lower-level features and relationships required as support. A composite pattern is paired to a set of target objects by pairing each component of its descriptor to a subset. For example, suppose the pattern descriptor for p in the example pair m above were $[p_1, p_2, p_3]$. Then p might be paired to $\{t_1, \ldots, t_n\}$ as follows: $[(p_1, \{t_1, t_2\}), (p_2, \{t_3\}), (p_3, \{t_4, \ldots, t_n\})]$.

Matches to an ideal pattern at any level will rarely be exact, due to mutations and various genetic rearrangements, and a differential measure of partial similarity is used to gauge overall plausibility. For example, the "spacer" primitive pattern allows for two flanking target objects to be separated by several amino acids. A separation slightly outside the allowable range (perhaps a genetic insertion) incurs a similarity score penalty. The larger the separation the larger the penalty, reflecting the biological intuition that long insertions are somewhat less likely than short ones. In our composite example, the similarity of p to $\{t_1, \ldots, t_n\}$, i.e. $\sigma(m)$, is computed recursively from $\sigma(p_1, \{t_1, t_2\})$, $\sigma(p_2, \{t_3\})$, and $\sigma(p_3, \{t_4, \ldots, t_n\})$.

8. PATTERN INVOCATION ALGORITHM.

Because a small number of patterns are hierarchically organized, the choice of *which rule* to invoke is usually unproblematic. For a number of reasons, however, perfect matches at any level are unlikely. The dominant problem becomes, not which rule to invoke, but to *which locations* in the protein the rule most plausibly applies. We map the rule invocation problem into a search problem, and search for groups which are sufficiently similar to the antecedent pattern.

The search for a differential similarity to a composite pattern consists of attempting to pair each component of its pattern descriptor to an admissible subset of target objects. A partial pairing, constructed at some intermediate stage, might pair only some of the descriptor components. For a given composite pattern, ARIADNE's search space is the set of all possible partial pairings. The single start node in this search space is the empty partial pairing, and goal nodes are complete pairings of all descriptor components. An operator which carries one partial pairing into its successors, is to expand the next unpaired

descriptor component by hypothesizing pairings to every admissible set of target objects. By applying this operator first to the start node and then iteratively to resulting partial pairings, all complete pairings may be found.

Since complete pairings are ordered by the similarity score σ and only the higher-scoring ones are of interest, an efficient search strategy is desirable. The well-known A* search [45] efficiently accommodates differentially inexact similarities to a descriptor and tends to focus search effort on the most promising candidates.[5] A* (see table 1) is a best-first branch-and-bound search with dynamic elimination of redundant pairings and an optimistic estimate of the contribution of the remaining unpaired descriptor components. (The elimination of redundant pairings may optionally be suppressed.) Optimality and convergence are both guaranteed.

The key to A* search is in the selection of which partial pairing to expand. Each partial pairing has a "best possible score", which is the highest score that the most favorable possible pairing of yet-unpaired descriptor components could ever yield. At each step the partial pairing with the highest best-possible-score is selected. If its best-possible-score is below the cut-off threshold the search can fail immediately, as *no* partial pairing could possibly exceed the threshold. Similarly, if it is a complete pairing then no other partial pairing can ever complete to a higher score. Otherwise, its next unpaired descriptor is expanded and the algorithm iterates. It is possible to enumerate all complete pairings in decreasing order of similarity score, pausing and continuing the search at will.

9. MONONUCLEOTIDE BINDING FOLD.

To illustrate the power of matching against secondary structure predictions we present a novel proposed protein alignment, found using this system, for the protein class of aminoacyl-tRNA synthetases. The proposed alignment agrees with the few existing alignments based on primary sequence similarities (or homologies) where such are known, and predicts novel alignments for some enzymes having no known primary sequence similarities. However, ARIADNE uses *no primary sequence similarities* in constructing this alignment.

The aminoacyl-tRNA synthetases help establish the rules of the genetic code, by mediating the translation of DNA to protein. They are responsible for attaching an amino acid to its corresponding tRNA, so that the tRNA can transfer the amino acid to a growing protein chain. It is known from co-crystal structures of the enzymes plus substrate that

[5] However, other problem areas with different underlying properties could exploit search strategies closer to that area's native structure. For example, hierarchical recognition in a well-structured problem area requiring only exact matches could employ a depth-first graph isomorphism search.

1. Form a queue of partial pairings, of pattern descriptor components to target object sets. Let the initial queue consist of the empty pairing having no descriptor components matched at all.

2. Until the queue is empty, a complete pairing is reached, or the upper-bound estimate of the best possible score falls below cutoff:

   2a. If the first pairing is complete, or its best possible score falls below cutoff, do nothing.
   2b. If the first pairing has some unpaired descriptor components:

      2b1. Remove the first pairing from the queue.
      2b2. Form new pairings from the removed pairing by matching its next unpaired descriptor component to possible groups of target objects.
      2b3. Add the new pairings to the queue.
      2b4. Sort the queue by an UPPER-BOUND estimate of the best similarity score which could be achieved by the most favorable possible pairing of the remaining unpaired components, highest-scoring pairings in front.
      2b5. If eliminating redundant pairings and two or more pairings pair the same pattern component to the same group of target objects, delete all those pairings except the one that has the highest similarity score for the objects paired to that point.

3. If a complete pairing has been found which is above cut-off threshold, announce success; otherwise announce failure.

Table 1. ARIADNE's A* Search Algorithm. This table has been adapted (by permission) from Winston [45].

the mononucleotide binding fold is involved in the binding of ATP and the amino acid [6].

The synthetases all bind similar substrates and all catalyze the same reaction [38], but have dissimilar primary sequences. A small region of six to eleven identical amino acids is known for four synthetases [43] (the tyrosyl-tRNA synthetase (TyrRS) from *B. stearothermophilus* and methionyl-tRNA synthetase (MetRS), isoleucyl-tRNA synthetase (IleRS), and glutamyl-tRNA synthetase (GlnRS) from *E. coli*), and this region is conserved in species variants. High-resolution X-ray crystal structures are available for only two of these enzymes (TyrRS and MetRS) [6, 50]. These two show a common super-secondary structure incorporating the small identical region: a 140 amino acid structure of nearly identical folding which includes the mononucleotide binding fold (see figure 3a). It has been of considerable interest to determine if this structure might also exist in the other synthetases, but the primary sequences are too dissimilar to support further inference.

We hypothesized a pattern descriptor for the synthetase mononucleotide binding fold [6] (see figures 3c and 6). This pattern (manuscript in press [44]) combines primary and secondary structure elements. It consists of three types of pattern elements (figure 6a): secondary structure objects (elements 1, 3, 5 and 7); amino acid objects (8 and 10); and spacer (or gap) objects (2, 4, 6 and 9). Secondary structures were hypothesized by the Chou and Fasman pseudoprobabilities [7, 34]. Spacer objects indicate the minimum and maximum number of amino acids between the flanking objects before a penalty is imposed. Object 5 is an α-helix with the "Gly+" dipeptide — a Glycine (G) amino acid immediately followed by an H-bond donor amino acid (C, K, H, N, Q, or R) — within four amino acids from the N-terminal (left) end of the α-helix.

This pattern was input to ARIADNE together with the protein data consisting of the fourteen primary sequences [3, 4, 14, 19, 21, 29, 30, 33, 39, 42, 43, 48, 49] annotated with secondary structure predictions of α-helices, β-sheet strands, and β-turns [7, 34]. A match was found once in twelve of the fourteen synthetases (figure 6b). The unique match in the *E. coli* MetRS and the *B. stearothermophilus* TyrRS corresponded well to the regions of the mononucleotide binding fold detected in the X-ray crystal structures. The unique matches in the Ile-, Tyr-, Met-, and Gln-tRNA synthetases include the small known region of identical amino acids. Species variants of the same synthetase often exhibit strong homologies, presumably indicating similar tertiary structure. The pattern was found in the region of the *E. coli* TyrRS homologous to the known structure from *B. stearothermophilus* TyrRs. The two matches for the tryptophanyl enzymes (TrpRS) were also found in homologous

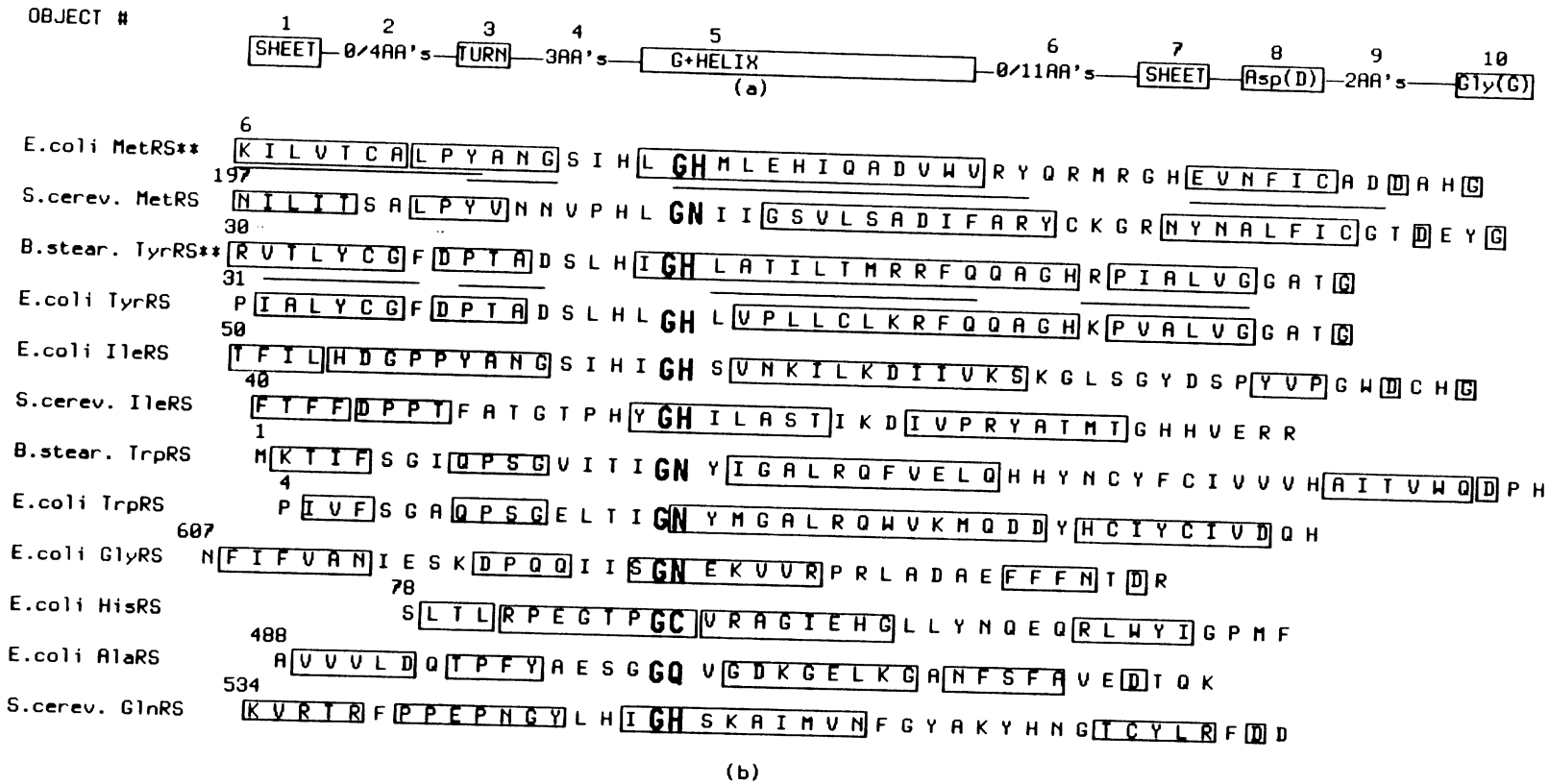


Figure 6. Proposed Mononucleotide Binding Fold alignment of aminoacyl-tRNA synthetases.

(a) Pattern for mononucleotide binding fold shown as a linear string of pattern objects (see text for description).
(b) Amino acid sequences (in the one-letter amino acid code) of the regions which matched the mononucleotide binding fold pattern with a positive similarity. The regions which correspond to objects 1, 3, 5, 7, 8 and 10 are boxed. The Gly and H-bond donor amino acid of object 5 (see text) are bold-faced. All boxed secondary structure objects were predicted by the ancillary program PRSTRC [34], based on Chou and Fasman pseudo-probabilities [7].
(**) The underlined amino acids are known by X-ray crystal data [5] to form the first sheet (object 1), the turn (object 3), the helix (object 5), and the second sheet (object 7) in the B. stear. TyrRS [6] and the E. coli MetRS [50].

Figure 6. The Proposed Mononucleotide Binding Fold Alignment.

regions. No matches were found in a set of fourteen structurally representative control enzymes known not to bind mononucleotides.

The aminoacyl-tRNA synthetase structure is the subject of ongoing research and analysis by molecular biologists. The example presented here is intended to illustrate the power and utility of the computational approach described in this paper, not to be a detailed biochemical analysis. Instead, it presents the most pedagogical of several patterns we have explored. To conform to the canons of science, a biochemical analysis would necessitate presentation of far more molecular biology than is appropriate for the intended audience of this paper. The technique would be calibrated against a functional group of proteins with known structure. A much larger set of control proteins would be analyzed in order to more conclusively characterize false-positive behavior. Other predictors would supplement the Chou-Fasman pseudoprobabilities. New synthetase sequences, published since this paper was written, would be analyzed. The pattern descriptor would be further refined to include all relevant biological knowledge. These tasks have been performed and largely corroborate the initial results given here, but their presentation is beyond the scope of this paper. Detailed analysis of the molecular biology aspects of the aminoacyl-tRNA synthetase mononucleotide binding fold functional domain will be published elsewhere [44].

10. DISCUSSION.

The principle sources of power in ARIADNE are:

1. The ability to entertain multiple, unreliable, inconsistent knowledge sources. Since no prediction scheme produces accurate predictions, any inference procedure which vitally depended on the consistency of its database (e.g., some forms of theorem-proving) would be ineffective.

2. The use of a pattern-similarity measure to guide flexible invocation of inference rules. This conveys a degree of robustness in the face of pattern fluctuations such as mutations.

3. Implementation of the rule-invocation similarity measure as an A* search [45]. This provides an efficient enumeration of match candidates, in order of decreasing similarity.

4. A flexible framework for pattern descriptor language development and extension. This is important because all the appropriate descriptor elements are surely not yet known.

5. Explicit identification and representation of the intermediate hierarchy, which helps in several ways:

   (a) Many of the higher-order (super-secondary) structures of interest are most effectively expressed in terms of lower and intermediate levels of hierarchy (secondary structure groupings), and not directly at the lowest level of description.

   (b) Handling patterns in small pieces encourages selective pattern refinement.

   (c) Expressing patterns consisting of key residues embedded in secondary structures involves the interaction of different hierarchical levels.

   (d) Breaking a large pattern into pieces increases search efficiency by reducing the potentially exponential time dependency on pattern size.

The approach presented here is limited to detecting similarities in patterns of known and/or predicted structural elements. To the extent that hypotheses of interest can be expressed in the form of a structural pattern, ARIADNE provides a powerful and efficient vehicle for finding supporting regions in the target proteins. However, no use is currently made of primary sequence similarities (or homologies), which would provide additional evidence for favoring some alignments over others. No direct use is made of three-dimensional spatial constraints (such as investigated by [20]). The secondary structure predictions remain inherently inaccurate, even though trade-offs can be made between reliability and coverage. No attempt has been made to encode or exploit "expert rule-of-thumb" knowledge of general biochemical heuristics.

Construction of abstract organizational hypotheses implies that low-level features meet the additional constraints imposed by higher-order patterns and relationships. These constraints take two forms: requiring a specified relationship with an element unambiguously present in the primary input (e.g., key amino acids); and requiring a specified relationship with other predicted or inferred features. Importantly, in a hierarchical pattern recognizer the structure imposed by higher-order patterns implies strong constraints on the admissibility and interpretation of low-level features, because those not fitting into a higher-level pattern will be dropped. A pattern acts to prune the (uncertain, heuristic, empirical) low-level features by selective attention, based on the strong constraint of fitting into higher-order organization (see figure 4a-e). Low-level features will be interpreted in terms of the expectations encoded in the patterns being searched for.

This has both good and bad aspects. When an intelligent agent (e.g., a biologist) hypothesizes and searches for the existence of a particular pattern based on supporting biochemical or circumstantial evidence, selective feature attention extends that evidential support down to low-level feature selection, and features supporting the pattern will be propagated upward. When a large number of patterns are sought randomly in a large number of targets (as in a database search), then each pattern will impose its own selective bias and additional confirming evidence should be sought. In either case, an important estimate of the false positive (resp. false negative) rate may be had by testing a control set known not to (resp. known to) actually satisfy the descriptor.

11. SUMMARY AND FUTURE RESEARCH.

We have described a flexible pattern-action framework for the recognition of molecular biological structures. The micro-world is characterized by recognizable higher orders of organization obscured by a high degree of uncertainty and imprecision, and the general approach should be applicable to similarly ill-structured problem areas. ARIADNE supports inexact but similar matches, direct representation of higher orders of organization, the use of ancillary secondary structure hypotheses, an extensible pattern-description language, and arbitrary actions on pattern invocation similar to a rule-based expert system. A novel proposed alignment of the aminoacyl-tRNA synthetases was found using this system.

We expect this to be useful for continuing research in the fields of both molecular biology and machine learning. Possible explorations in molecular biology include further investigation of patterns suspected to exist at the super-secondary level, as well as alternate independent sources of the low-level feature hypotheses. Possible explorations in machine learning include use of this system as an hypothesis verification mechanism for some other system which proposes hypothesized similarities. We are exploring automatic pattern discovery based on empirical regularities, but results are too preliminary to discuss here.

ACKNOWLEDGMENTS.

Thanks to Paul Schimmel and Patrick Winston for advice and encouragement, and to Russ Altman, Diane Apostolakos, John Batali, Michael Brent, David Chapman, John Mallery, Mira Marcus, Dave McAllester, Bob Rogers, David Saslav, Eric Saund, and Dan Weld for discussion and assistance. Personal support for the first author was furnished by an IBM Graduate Fellowship, and during the early stages of this research was furnished by an NSF Graduate Fellowship and by a research/teaching assistantship from MIT. This

paper describes research performed jointly at MIT's Artificial Intelligence Laboratory, and at Harvard Medical School's Molecular Biology Computer Research Resource in the Dana Farber Cancer Institute. Support for the MIT Artificial Intelligence Laboratory's research is provided in part by ONR contract N00014-80-C-505. Support for the Molecular Biology Computer Research Resource's research is provided in part by NIH grant number RR02275.

REFERENCES.

[1] Abarbanel, R. M. *Protein Structural Knowledge Engineering.* Ph.D. thesis, U. Cal. San Francisco, 1984.

[2] Anfinsen, C. B., Haber, E., Sea, M., and White, F. H. *Proc. Natl. Acad. Sci. USA* 47 (1961), 1309-1314.

[3] Barker, D. G., Bruton, C. J., and Winter, G. *FEBS Lett.* 150 (1982), 419-423.

[4] Barker, D. G., Ebel, J.-P., Jakes, R., and Bruton, C. J. *Eur. J. Biochem.* 127 (1982), 449-457.

[5] Blow, D. M., Bhat, T. N., Matcalfe, A., Risler, J. L., and Brunie, S. *J. Mol. Biol.* 171 (1983), 571-576 (Table 1).

[6] Blow, D. M., and Brick, P. *Biological Macromolecules and Assemblies 2: Nucleic Acids and Interactive Proteins.* John Wiley and Sons, New York, 1985.

[7] Chou, P. Y. and Fasman, G. D. *Biochemistry* 13 (1974), 222-245.

[8] Cohen, F. E., Richmond, T. J., and Richards, F. M. *J. Mol. Biol.* 132 (1979), 275-288.

[9] Cohen, F. E., et al. Turn Prediction In Proteins: A Complex Pattern Matching Approach. *Biochemistry* 25 (1986), 266-275.

[10] Creighton, T. *Protein Structure and Molecular Properties.* W. H. and Company, New York, 1983.

[11] Davis, R., and Lenat, D. *Knowledge-Based Systems in Artificial Intelligence.* McGraw-Hill, New York, 1982.

[12] Doolittle, R. F., Hunkapillar, M. W., Hood, L. E., Davare, S. G., Robbins, K. C., Aaronson, S. A., and Antoniades, H. N. *Science* 221 (1983), 275-277.

[13] Eisenberg, D., Weiss, R. M., Terwilliger, T. C. The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity. *Proc. Natl. Acad. Sci. USA. Biophysics* 81 (Jan. 1984) 140-144.

[14] Freedman, F., Gibson, B., Donovan, D., Biemann, K., Eisenbeis, S., Parker, J., and Schimmel, P. *J. Biol. Chem.* 260 (1985), 10063-10068.

[15] Friedland, P., and Iwaskai, Y. The Concept and Implementation of Skeletal Plans. *J.*

of Automated Reasoning 1:2 (1985), 161-208.

[16] Friedland, P., and Kedes, L. Discovering the Secrets of DNA. *COMPUTER* 18:11 (Nov. 1985), 49-69.

[17] Garnier, J., Osguthorpe, D. J., and Robsen, B. *J. Mol. Biol.* 120 (1978), 97-120.

[18] Gascuel, O., and Danchin, A. To appear, *J. Mol. Evol.* (1987).

[19] Hall, C. V., vanCleemput, M., Muench, K. H., and Yanofsky, C. *J. Biol. Chem.* 257 (1982), 6132-6136.

[20] Hayes-Roth, B., et al. PROTEAN: Deriving Protein Structure From Constraints. In *Proc. AAAI Fifth Natl. Conf. on Artificial Intelligence* (Philadelphia, Penn., Aug. 11-15, 1986). Morgan Kaufman, Los Altos, Ca., 1986, pp. 904-909.

[21] Hoben, P., Royal, N., Cheung, A., Yamao, F., Beimann, K., and Soll, D. *J. Biol. Chem.* 257 (1982), 11644-11650.

[22] Kolata, G. Trying to Crack the Second Half of the Genetic Code. *Science* 233:4768 (5 Sept. 1986), 1037-1040.

[23] Kyte, J., and Doolittle, R. *J. Biol. Mol.* 157 (1982), 105-132.

[24] Lathrop, R. H., and Kirk, R. S. A System Which Uses Examples To Learn VLSI Structure Manipulations. In *Proc. AAAI Fifth Natl. Conf. on Artificial Intelligence* (Philadelphia, Penn., Aug. 11-15, 1986). Morgan Kaufman, Los Altos, Ca., 1986, pp. 1024-1028.

[25] Lesk, A. M. *J. Biol. Mol.* 136 (1980), 225-270.

[26] Levitt, M. *J. Mol. Biol.* 170 (1983), 723-764.

[27] Lim, V. I. *J. Mol. Biol.* 88 (1974), 857-894.

[28] Lindsay, R., et al. *DENDRAL.* McGraw-Hill, New York, 1980.

[29] Ludmerer, S., Ph.D. thesis, Massachusetts Institute of Technology, 1986.

[30] Mayaux, J. F., Fayat, G., Frommant, M., Springer, M., Grunberg-Manago, M., and Blanquet, S. *Proc. Natl. Acad. Sci. USA* 80 (1983), 6152-6156.

[31] Minsky, M. *Society of Mind.* Simon and Schuster, New York, to appear, 1987.

[32] Ptitsyn, O. B., and Rashin, A. A. *Biophys. Chem.* 3 (1975), 1-20.

[33] Putney, S. D., Royal, N. J., de Vegar, H. N., Herlihy, W. C., Biemann, K., and Schimmel, P. R. *Science* 213 (1981), 1497-1501.

[34] Ralph, W., Webster, T. A., and Smith, T. A Modified Chou and Fasman Protein Structure Algorithm. Submitted for publication; *PRSTRC* program available from MBCRR, Dana Farber Cancer Institute, 44 Binney St., Boston, Mass.

[35] Rich, C. Inspection Methods in Programming. M.I.T. Artificial Intelligence Laboratory Technical Report 604, June 1981.

[36] Richardson, J. *Advances in Protein Chemistry* 34 (1981), 167-339.

[37] Sanoff, D., and Kruskal, J. B. (eds.) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* Addison-Wesley, Reading, Mass., 1983.

[38] Soll, D., and Schimmel, P. R. *Ann. Rev. Biochem.* 48 (1979), 601-648.

[39] Walter, P., Gangloff, J., Bonnet, J., Boulanger, Y., Ebel, J.-P., and Fasiolo, F. *Proc. Natl. Acad. Sci. USA* 80 (1983), 2437-2441.

[40] Waterman, D. and Hayes-Roth, F. (eds.) *Pattern-Directed Inference Systems.* Academic Press (New York), 1978.

[41] Waterman, M. S. *Bull. of Math. Biol.* 46 (1984), 473-500.

[42] Webster, T. A., Gibson, B. W., Keng, T., Biemann, K., and Schimmel, P. *J. Biol. Chem.* 258 (1983), 10637-10641.

[43] Webster, T. A., Tsai, H., Kula, M., Mackie, G., and Schimmel, P., *Science* 226 (1984), 1315-1317.

[44] Webster, T. A., Lathrop, R. H., and Smith, T. F. Prediction of a Common Structural Domain in Aminoacyl-tRNA Synthetases Through Use of a New Pattern-Directed Inference System, to appear, *Biochemistry.*

[45] Winston, P. H. *Artificial Intelligence, 2nd ed.* Addison-Wesley, Reading, Mass., 1984, 113-114.

[46] Winston, P. H. Learning and Reasoning by Analogy. *Comm. ACM* 23:12 (December, 1980), 689-703.

[47] Winston, P. H., Binford, T. O., Katz, B., Lowry, M. Learning Physical Descriptions from Functional Descriptions, Examples, and Precedents. In *Proc. of the Natl. Conf. on Artificial Intelligence* (Washington, D. C., Aug. 22-26, 1983). William Kaufman, Los Altos, Ca., 1983, pp. 433-439.

[48] Winter, G. P., and Hartley, B. S. *FEBS Lett.* 80 (1977), 340-342.

[49] Winter, G., Koch, G. L. E., Hartley, B. S., and Barker, D. G. *Eur. J. Biochem.* 132 (1983), 383-387.

[50] Zelwer, C., Risler, J. L., Brunie, S. *J. Mol. Biol.* 155 (1982), 63-81.

CS-TR Scanning Project
Document Control Form

Date : 10/18/95

Report # AIM-902

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 24 (28-IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☐ DOD Form ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages(by page number):_____

Photographs/Tonal Material (by page number):_____

Other (note description/page number):

Description :	Page Number:
IMAGE MAP: (1-24) UNNUMBERED TITLE PAGE, 1-23	
(25-28) SCANCONTROL TRGT'S (3)	

Scanning Agent Signoff:

Date Received: 10/18/95 Date Scanned: 10/24/95 Date Returned: 10/26/95

Scanning Agent Signature: Michael N. Cook

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

Scanned

Date: 10/24/1995

M.I.T. Libraries
Document Services