Machine Learning, 21, 151–175 (1995)
© 1995 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Searching for Representations to Improve Protein Sequence Fold-Class Prediction

THOMAS R. IOERGER ioerger@cs.uiuc.edu
National Center for Supercomputing Applications, The Beckman Institute
Department of Computer Science, University of Illinois, Urbana, IL 61801

LARRY A. RENDELL rendell@cs.uiuc.edu
National Center for Supercomputing Applications, The Beckman Institute
Department of Computer Science, University of Illinois, Urbana, IL 61801

SHANKAR SUBRAMANIAM shankar@ncsa.uiuc.edu
National Center for Supercomputing Applications, The Beckman Institute
Department of Physiology and Biophysics, University of Illinois, Urbana, IL 61801

Editors: Jude Shavlik, Lawrence Hunter, and David Searls

Abstract. Predicting the fold, or approximate 3D structure, of a protein from its amino acid sequence is an important problem in biology. The homology modeling approach uses a protein database to identify fold-class relationships by sequence similarity. The main limitation of this method is that some proteins with similar structures appear to have very different sequences, which we call the "hidden-homology problem." As in other real-world domains for machine learning, this difficulty may be caused by a low-level representation. Learning in such domains can be improved by using domain knowledge to search for representations that better match the inductive bias of a preferred algorithm. In this domain, knowledge of amino acid properties can be used to construct higher-level representations of protein sequences. In one experiment using a 179-protein data set, the accuracy of fold-class prediction was increased from 77.7% to 81.0%. The search results are analyzed to refine the grouping of small residues suggested by Dayhoff. Finally, an extension to the representation incorporates sequential context directly into the representation, which can express finer relationships among the amino acids. The methods developed in this domain are generalized into a framework that suggests several systematic roles for domain knowledge in machine learning. Knowledge may define both a space of alternative representations, as well as a strategy for searching this space. The search results may be summarized to extract feedback for revising the domain knowledge.

Keywords: domain knowledge, change of representation, theory revision, protein structure prediction, homology modeling, amino acid properties

1. Introduction

Studies of learning methods applied to real-world domains have contributed a great deal to the field of machine learning. In this paper, we present a machine learning approach to the problem of protein structure prediction. The results of this inter-disciplinary research contribute directly to the field of protein science. In addition, we will draw some generalizations for machine learning, based on the techniques we developed for this domain.

Protein structure prediction is an important problem in biology. Proteins are encoded by genetic sequences in the DNA of a cell. When a protein is synthesized in the cell, it folds into a complex, three-dimensional shape, or tertiary structure, which determines the biological function of the protein. In recent years there has been an increase in the rate at which the sequences of proteins are being determined (Watson, 1990). However, the structures of proteins are much harder to determine in a laboratory. To analyze the large volume of sequence data available, and to gain the deepest insights into the biological functions of various proteins, we need to develop new computational methods for predicting the structures of proteins from their sequences.

One of the most successful approaches for global structure prediction to date has been to use *homology modeling* (Blundell et al., 1987). Homology modeling refers to the process of aligning the sequence of a protein whose structure is unknown to the sequence of a protein whose structure is known. If the quality of the sequence alignment (degree of match) is significantly high, then it may be assumed that the two proteins have similar structures. The use of homology modeling to predict the structure of a protein requires a protein database to provide sequences with which to compare.

Protein structures fall into fairly distinct clusters of overall shape called *fold classes* (Pascarella & Argos, 1992). A database of protein structures can be thought of as a set of examples of known folds. Rather than actually returning the atomic coordinates for a matched protein as a structure approximation, it is sufficiently informative to identify the fold class of a protein (Subramaniam et al., 1992).

Homology modeling is limited because some proteins that have the same fold do not have any apparent sequence similarity, which we call the "hidden-homology problem." Consequently, a set of proteins with similar structures may not constitute a sufficient representation of the fold. This results in under-utilization of protein databases because some potential fold relationships are not being recognized. Although it is unclear how many relationships among proteins are currently undetected, hidden homology may explain why there are so many sequences whose folds are unknown.

In this paper, we treat fold-class prediction from protein sequences as a machine learning problem. The use of homology modeling with a protein database can be thought of as a domain-specific learning method in which the database serves as a set of training examples. The hidden-homology problem causes poor learning performance in terms of accuracy. Our general goal is to use advanced machine techniques to improve the predictive accuracy in this domain.

We suggest that protein fold-class prediction is a difficult domain for machine learning because the initial representation of protein sequences is "low-level." However, a considerable amount of knowledge is available regarding the physical and chemical properties of the constituents of protein sequences, called amino acids. While biophysicists are far from having a comprehensive theory of protein folding, they have acquired some general rules for the roles amino acids can play in determining protein structure (Dayhoff et al., 1972; Richardson & Richardson, 1989).

We take a systematic approach to improving the learning performance in this domain by applying this biophysical knowledge to change the *representation* of examples. We will show how various properties of amino acids can be expressed in a formalism based on partitions of the amino acids. Through the partition formalism, domain knowledge can be used to suggest ways to re-represent protein sequences in terms of amino acid properties, rather than identities.

A partition can capture many possible relationships among the amino acids based on various combinations of properties. Thus we propose a *search framework* for improving learning in this domain, in which the partitions define a search space. Each such partition may be evaluated by using it to re-represent a set of example sequences whose fold classifications are known, and then computing the effect on the accuracy of homology modeling within that set. The goal is to identify partitions that, when used to re-represent protein sequences, increase the accuracy of homology modeling.

We present some results from a series of experiments using a particular set of example proteins to test the method for change of representation we have proposed in this domain. We show that this method can improve the accuracy of fold classification by discovering better representations for amino acid sequences. We also demonstrate how the search results can be summarized to extract refinements of the domain knowledge.

Finally, we propose a major extension to the representation of amino acid sequences that takes sequential context into account. Biophysical knowledge suggests that there are multiple dimensions of similarity among the amino acids due to combinations of chemical and physical properties (Richardson & Richardson, 1989). While the initial partition formalism is limited in its ability to express multiple relationships, the roles an amino acid might be playing at a given site might be context-dependent (Overington et al., 1992). We present some initial results from incorporating properties of neighboring residues into the representation of each amino acid in a protein sequence to explore more subtle relationships among the amino acids suggested by our earlier results.

The methods we develop to improve the accuracy of prediction in this domain can be generalized to improve machine learning in other difficult domains as well. Specifically, the framework we establish for searching for alternative representations can be applied to other real-world domains, especially where domain knowledge is available. This framework provides specific roles for knowledge to facilitate learning. Domain knowledge can define a space of alternative representations, and might be used to identify a search strategy for focusing on regions of the space that are likely to contain representations that improve predictive accuracy. The search can be guided by incremental improvements in the performance of the inductive algorithm as it is applied to a set of training example in various representations.

One advantage of our search framework is that it not only specifies how knowledge may be incorporated into learning, but it also suggests how refinements of that knowledge might be extracted from the search results. The results produced by the search process can be analyzed to provide feedback for critiquing the initial knowledge based on utility. Thus this method for change of representation can also be seen as a semi-automated or interactive form of theory revision. In a real-world domain, such feedback can be as significant as gains in predictive accuracy.

Table 1. The 20 amino acids and their three-letter and one-letter symbols.

alanine	Ala	A	isoleucine	Ile	I	arginine	Arg	R
cysteine	Cys	C	lysine	Lys	K	serine	Ser	S
aspartate	Asp	D	leucine	Leu	L	threonine	Thr	T
glutamate	Glu	E	methionine	Met	M	valine	Val	V
phenylalanine	Phe	F	asparagine	Asn	N	tryptophan	Trp	W
glycine	Gly	G	proline	Pro	P	tyrosine	Tyr	Y
histidine	His	H	glutamine	Gln	Q			

2. Overview of Protein Structures

Proteins are macromolecules produced by cells and used for a wide variety of biological functions (Stryer, 1988). These macromolecules are linear polymers of smaller compounds called *amino acids* or *residues*. There are 20 standard amino acids, each having a common part that contributes to the backbone of the protein, plus a unique chemical group called a side-chain. The names and common abbreviations of the amino acids are listed in Table 1 for convenience. Each protein consists of a unique sequence of amino acids and is typically 100 to 500 residues in length.

There are several steps in the protein synthesis process (Stryer, 1988), the final stage of which is folding. Folding can take from milliseconds to several minutes, although the actual pathways are not well understood (Baldwin, 1989). Most proteins—at least soluble ones—fold into unique, globular (roughly spherical) shapes. White (1961) demonstrated that, after denaturation, proteins can re-fold into their original shapes in the absence of all other cellular components, which shows that the amino acid sequence alone is sufficient to determine the structure of a protein.

Some of the forces that drive protein folding (see Dill, 1990) include enthalpies of non-covalent interactions (e.g. hydrogen bonding) and entropic effects related to solvation (e.g. the hydrophobic effect). Proteins do undergo shape changes due to thermal vibrations, as well as during substrate binding, but these dynamic effects span time and length scales that do not affect the global topology of proteins in their native state (McCammon & Harvey, 1987).

The structure of a protein can be described at various levels of detail. The *primary structure* of a protein refers to its sequence of amino acids. *Secondary structure* is defined by the local conformations of each residue with respect to its neighbors (i.e. torsion angles for covalent bonds in the backbone, Schulz & Schirmer, 1979). Contiguous stretches of amino acids often form identifiable sub-structures called α-helices and β-sheets, connected by various kinds of loops and turns (Richardson, 1981). The global arrangement of these sub-structures is called the *tertiary structure*. Thus the globular shape into which a protein folds, which is precisely given by atomic coordinates, has an internal structure that is approximately specified by the path of the backbone.

By studying the qualitative patterns among protein tertiary structures, Richardson (1981) identified a taxonomy of protein *folds*. Many of the proteins whose structures have been determined appear to cluster into groups that have the same overall shape, defining fold classes (Pascarella & Argos, 1992). Proteins of a given fold often have similar functions and, where there is evidence that they are evolutionarily related, appear to have diverged from a common ancestor (Doolittle, 1981). However, other folds contain members that seem to be adapted for distinct purposes (Chothia, 1988), suggesting that they might have converged to one of a restricted set of allowable folds (Finkelstein & Ptitsyn, 1987). Based on the statistics of database growth, Chothia (1992) has estimated that only around 1000 folds are used in all biological systems, of which more than 100 are already known.

The amino acid sequences of proteins are now routinely determined by DNA sequencing. However, the physical determination of protein tertiary structures is much more difficult. The two most common methods, X-ray crystallography and NMR spectroscopy, both require complex laboratory preparation, as well as computationally-intensive data analysis. As a consequence, tens of thousands of protein sequences are known, but the structures of only a few hundred proteins have been solved so far.

Thus predictive methods are needed to compute the structure of a protein from its sequence. Several algorithms exist for predicting the local secondary structure of sites within a protein (Chou & Fasman, 1974; King & Sternberg, 1990; Qian & Sejnowski, 1988). However, the results of these methods have generally not been assembled into global predictions of tertiary structure. Molecular dynamics simulations have been used to find atomic configurations of minimum energy (McCammon & Harvey, 1987). But a *de novo* simulation of folding is computationally intractable (Richards, 1992), so these methods are more often used for refining approximately-correct structures. In the next section, we discuss homology modeling as one of the most successful methods to date for predicting protein tertiary structure.

3. Homology Modeling

Homology modeling can be used to predict the structure of an unknown protein by aligning it with example proteins whose structures are known (Blundell et al., 1987). If the sequence of interest matches a known protein, it may be assumed to be in the same fold class, identifying its approximate structure. Thus, unlike molecular simulation of folding from first principles, homology modeling exploits databases of known structures, such as the Brookhaven Protein Databank (PDB, Bernstein et al., 1977), as examples of the complex relationship between protein sequence and structure.

Because approximate structures are still highly informative of a protein's function, and can be further refined by energy minimization (Nell et al., 1992), homology modeling reduces the goal of structure prediction to the task of fold-class identification (Subramaniam et al., 1992). This goal is much less complex than predicting atomic coordinates because there are many fewer degrees of freedom. Protein databases can be thought of as a set of examples for fold classes: the examples are described by amino acid sequences, and each sequence is labeled by its known fold class.

The use of sequence alignment in homology modeling corresponds to an indexing mechanism for a protein database. A sequence whose structure is to be predicted is aligned with each example. If a significant match is found, its fold classification is looked up and returned as the structure prediction for the query protein. Although this method cannot predict the structure of a protein for which there are no examples of the same fold in the database, the estimation of around 1000 total fold classes implies that protein databases should eventually converge upon a complete library (Chothia, 1992).

3.1. Definitions of Sequence and Structure Similarity

Sequence similarity is based on an alignment score. A sequence alignment establishes a correspondence between the amino acid residues in each sequence that preserves order but allows gaps. The gaps are assigned costs relative to the value of mismatches between residues (Gotoh, 1982; Fitch & Smith, 1983). Various algorithms exist to find an alignment of maximum score, which represents a globally optimal balance between the costs of amino acid mismatches and the costs of gaps (Needleman & Wunsch, 1970; Smith & Waterman, 1981).

Alignment scores are relevant to protein tertiary structure prediction because they reflect evolutionary processes that cause protein structures to diverge. When two proteins evolve from a common ancestral gene, they accumulate independent mutations that distinguish their sequences. The most frequent kind of mutation is the replacement of one amino acid by another; insertions or deletions of amino acids can also occur (Stryer, 1988). The alignment score counts amino acid replacements through residue mismatch penalties, and it counts insertions and deletions through gap penalties. The accumulation of such differences in the amino acid sequences of two proteins contributes incrementally to the differences in their structures (Chothia & Lesk, 1986).

A single alignment score by itself is not very informative because any two sequences can be aligned, generating some arbitrary score. For example, two sequences drawn randomly with replacement from an alphabet of 20 symbols could be expected to have around 5% similarity on average (without gaps). A common way of estimating the significance of a given alignment score is to compare it to a distribution of alignment scores between actual proteins with different structures (Doolittle, 1981). Proteins in different fold classes can have sequence alignment scores as high as 25% because of uneven frequencies of occurrence of the amino acids in biological systems (Sander & Schneider, 1991). If the alignment of two sequences produces a score that is three or more standard deviations above this distribution, they can be assumed to belong to the same fold class (Lipman & Pearson, 1985).

True structural similarity is formally measured by RMSD, the root mean square deviation of the positions of corresponding backbone atoms in three-dimensional space (McLachlan, 1972). To compare the structures of two proteins, they must first be rotated relative to each other so that their internal structures overlap as well as possible. Then residues are assigned correspondences between the two structures, like a three-dimensional version of sequence alignment, and insertions, which often cause gaps in loop regions at the surfaces of a protein, are clipped out. Each amino acid in a protein has a reference point called the C_α atom; distances between these atoms in corresponding residues of the aligned structures are used to compute the RMSD.

Although it is more difficult to estimate the significance of a structural similarity score, observations of known protein structures suggest that two proteins with an RMSD value of less than around 3.0Å belong to the same fold class. Pascarella and Argos (1992) used such criteria to cluster 254 of the protein structures in the PDB into 83 distinct classes, including 38 folds with multiple examples.

Chothia and Lesk (1986) have demonstrated that the measures of sequence and structure similarity are correlated. They found that RMSD fits an exponential function of the alignment score over a wide variety of proteins. This observation establishes the principle that proteins with similar amino acid sequences have similar structures, which is the basis for homology modeling.

3.2. The Hidden-Homology Problem

The primary limitation of homology modeling is that, while proteins with similar sequences always have similar structures, the converse is not true. Proteins with similar structures do not always have significantly high alignment scores. For example, mandelate racemase and muconate lactonizing enzyme have only 26% amino acid identity in their sequences, which was not significant enough to indicate any relationship *a priori.* But when their structures were solved, they were observed to have an RMSD of only 1.3Å, and both belong to the β-barrel fold class (Neidhart et al., 1990).

We call this limitation the "hidden-homology problem." Hidden homology can cause some fold relationships to not be detected during homology modeling with a database. A particular protein might indeed belong to a known fold class, but its alignment to all examples could be too low to signal a match. Figure 1 illustrates the effect that hidden homology can have on the use of homology modeling for fold-class prediction. The consequence of this limitation of homology modeling is that protein databases are likely being under-utilized.

Although it is unclear how pervasive this problem is, there are many sequences whose structures remain unknown. Hidden homology implies that these proteins cannot reliably be excluded from known folds just because sequence similarity with known examples is low. Doolittle (1986) has called sequence alignment scores less than around 25% the "twilight zone" because such scores can neither be used to include nor exclude a sequence from a given fold class.

A variety of approaches have been taken to extend homology modeling to cases where structurally similar proteins have low sequence similarity. One approach has been to use partial match scores in alignments to reward mismatches that are more commonly observed as substitutions between related proteins (Schwartz & Dayhoff, 1978). Another approach has been to thread the sequence of the unknown protein into the structure of the known protein and evaluate the fit by various criteria (Jones et al., 1992). Multiple examples of a fold class may be used to construct a statistical profile with which alignments can be made (Gribskov et al., 1988). Finally, symbolic approaches such as ARIADNE (Lathrop et al., 1987) and PIMA (Smith & Smith, 1990) attempt to derive a consensus pattern from multiple sequences.

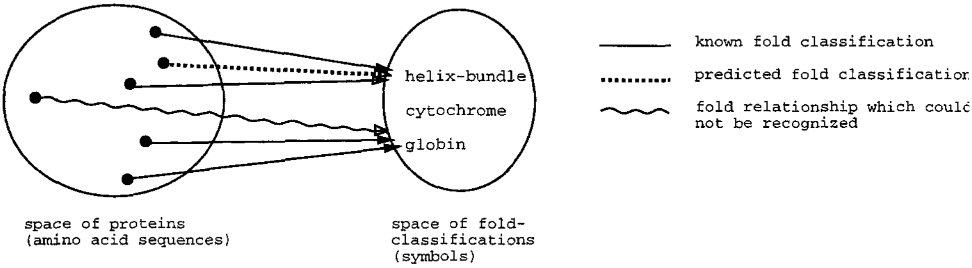
helix-bundle
cytochrome
globin
known fold classification
predicted fold classificatior
fold relationship which coulc not be recognized
space of proteins (amino acid sequences)
space of fold-classifications (symbols)

Figure 1. The effect of the hidden-homology problem on fold-class prediction. Using homology modeling, the fold of one of the helix-bundle proteins is predicted to be the same as that for its nearest neighbor (dotted line). However, hidden homology causes a failure to recognize one of the members of the globin fold (wavy line). Although this protein belongs to the class, the relationship cannot be recognized because the protein's sequence is so different from the other examples of this fold, and is nearly equally as distant to members of other folds.

Table 2. Components of an abstract model of machine learning, their instantiation for protein fold-class prediction, and examples from this domain.

Component	Instantiation	Examples
objects	proteins	human alpha hemoglobin, bovine serum albumin...
descriptions	amino acid sequences	Met-Ala-Glu-Leu-...
classifications	folds	helix-bundle, cytochrome, Ca-binding domain...
inductive algorithm	nearest-neighbor	homology modeling

4. Protein Fold-Class Prediction as a Machine Learning Problem

Protein fold-class prediction can be treated as a machine learning problem (Subramaniam et al., 1992). Machine learning problems often consist of a space of objects described in some language, and a classification of those objects into subsets (Michalski, 1983). Typically the goal is to use an inductive algorithm to draw generalizations from a set of pre-classified training examples so that classification of unseen test cases is more accurate than guessing.

Table 2 shows how these abstract components of machine learning are instantiated in the fold-class prediction domain. The objects in this domain are proteins, the space of descriptions consists of amino acid sequences, and fold classes divide the instance space into subsets. Homology modeling can be thought of as the inductive algorithm for protein fold-class prediction, with the protein database acting as a set of training examples. Because examples are explicitly saved for comparison during prediction tasks, homology modeling is a kind of nearest-neighbor, instance-based learning (Aha et al., 1991). The sequence alignment score between pairs of instances serves as the distance metric in this space.

Table 3. Major exchange groups of amino acids, based on the pairwise substitution frequency data collected by Dayhoff et al. (1972), and the properties they identify. For the convenience of this paper, abbreviations of the properties are also given.

amino acids	property	abbrev.
MET,ILE,LEU,VAL	large-and-hydrophobic	hyp
PHE,TYR,TRP	aromatic	aro
LYS,HIS,ARG	positive	pos
ALA,CYS,ASP,GLU,GLY,ASN,PRO,GLN,SER,THR	small	sml

We suggest that protein fold-class prediction is difficult for machine learning specifically because the initial representation in this domain is too low-level with respect to the algorithmic bias (Mitchell, 1980). The amino acid identities in protein sequences are too detailed, causing homology modeling to fail to recognize fold relationships.

However, an important technique in machine learning is to change the representation to better match the algorithmic bias (Michalski, 1983; Utgoff, 1986). One specific method for change of representation is feature construction (Matheus, 1989). Feature construction cannot be applied directly to the protein structure-prediction domain because examples are not represented as fixed-length feature-vectors. The variable-length sequences do not allow features to be uniquely identified in terms of sequence elements. Instead, our approach is to uniformly change the set of symbols used in protein sequences to reflect amino acid properties.

5. Change of Representation for Protein Fold-Class Prediction

The identities of amino acids in a protein sequence are over-specific for fold-class prediction; evolutionarily-related proteins have the same fold, but differ in their primary sequences at various positions. However, it has been observed that amino acid differences tend to fall into patterns according to certain physical and chemical properties.

These patterns have been quantified by Dayhoff et al. (1972), who analyzed substitution frequencies between pairs of amino acids found at the same positions in evolutionarily-related proteins. The substitution frequencies revealed several prominent subsets of amino acids called *exchange groups*, listed in Table 3. The distribution of amino acids found at any given site among similar proteins typically falls within only one of the exchange groups.

The exchange groups can be interpreted according to the notion that members within a group share a set of chemical and physical properties that are not common to members of other groups (see Table 3). These properties are relevant to protein structure prediction because they participate in the local structure around each residue. For example, some residues have polar side-chains that can form hydrogen-bonds. Other residues are small, so they can pack into crowded pockets.

Taylor (1986) proposed an extended list of eight properties—aliphatic, hydrophobic, aromatic, polar, charged, positive, small, and tiny—that define interesting subsets of amino acids. Based on our knowledge of biophysics, any of these properties could potentially be playing a structural role at a given position in the three-dimensional structure. Richardson and Richardson (1989) have given detailed descriptions of the amino acids in terms of properties such as constraint on backbone flexibility, degrees of freedom in the side-chain, occurrence of particular chemical subgroups, and frequency with which these amino acids participate in or alter certain secondary structures. Factor analyses of 163 properties of amino acids have demonstrated the primary importance of bulk and hydrophobicity (Kidera et al., 1985).

The relationships among the amino acids, defined by their properties, have often been used in *ad hoc* ways to justify alignments of protein sequences in regions where there is little amino acid identity. The properties mentioned above have also been used to generalize a set of sequences that belong to the same fold class (Lathrop et al., 1987; Smith & Smith, 1990). However, these properties have rarely been used to systematically re-represent individual protein sequences to improve the accuracy of fold-class prediction.

To formalize this domain knowledge in a way that can be applied to the representation of protein sequences, we initially propose the construction of a partition of the 20 amino acids. A set of mutually exclusive subsets suggested by Dayhoff's exchange groups is {{MILV}{HRK}{FWY}{ACDEGNPQST}}, which we call the "Dayhoff partition." This partition simultaneously expresses the properties listed in Table 3, and their abbreviations can be used as class names.

A partition can be used to re-represent a protein sequence by *transforming* each amino acid in the sequence into its class name. For example, a sequence such as Met-Lys-Ala... would become hyp-pos-sml... after the Dayhoff partition is applied.

Such a re-representation can have a significant effect on homology modeling. If pairs of sequence are re-aligned after being transformed, many more local matches will occur, even between proteins with different structures. However, proteins that belong to the same fold class must have similar sequences of properties in order to fold into similar structures. Because their amino acids are constrained to fulfill similar roles, this new representation may produce an additional increase in local matching between proteins in the same fold class.

Thus this method for re-representing protein sequences better matches the bias in the alignment algorithm and could facilitate the recognition of fold-class relationships. The Dayhoff partition seems be a good intermediate-level representation, exposing features relevant to protein structure (amino acid properties) while masking insignificant details (amino acid identities).

However, there is some uncertainty in the background knowledge in this domain about which properties are most important. Dayhoff et al. (1972) remarked that classifications of some amino acids is difficult because it is unclear to which group they most belong. This phenomenon is caused by the fact that amino acids have multiple properties, defining orthogonal dimensions of similarity. For example, threonine is like serine because they both contain a hydroxyl group, but it is like valine in shape.

This uncertainty in domain knowledge suggests a whole space of possible representations. The partition formalism can be used to express many alternative group definitions.

For example, if a given partition classifies threonine as a polar residue, then there exists a variant of this partition in which threonine occurs in the class containing valine.

Because any of these partitions may be applied to transform a protein sequence, the set of all possible partitions of the 20 amino acids defines the space of representations that are alternatives to using amino acid identities. In the following section, we show how to search this space to find representations that improve the accuracy of fold-class prediction.

However, the partition formalism is ultimately limited in expressive power; because groups of amino acids cannot overlap, a partition cannot capture the multiple dimensions of similarity. In fact, even Dayhoff's tables of pairwise substitution frequencies average together the effects of all the various roles amino acids can play. Domain knowledge suggests that these effects might be context-dependent, since the properties of an amino acid generally interact with neighboring residues to define the local environment. In section 6.3, we will show how this knowledge can be used to extend the partition formalism to incorporate context information directly into the representation.

6. Experiments and Analysis

In this section, we apply several search techniques to explore the space of representations constructed from partitions of amino acids. We take the Dayhoff partition as a starting node since domain knowledge suggested that it expresses relevant amino acid properties. First we compare this representation to the original amino acid identities, which can be thought of as a "null" partition with 20 singleton classes.

To evaluate these two partitions, we used a data set consisting of the 179 proteins shown in Table 4, classified into 35 folds by Pascarella and Argos (1992). The sequences themselves were taken from the Brookhaven Protein Databank (PDB, Bernstein et al., 1977), and sub-sequences and/or chains were extracted where specific domains were required.

The accuracy of homology modeling was estimated within this data set by computing alignment scores for each pair of sequences. For these experiments, we define the accuracy of homology modeling to be the frequency with which the sequence most similar to each sequence is in the same fold class.

We implemented a space-efficient, quadratic-time sequence-alignment algorithm by Myers and Miller (1988) on a Connection Machine (CM-5) to do all 15,931 pairwise alignments within our data set in parallel. Based on the recommendations of Fitch and Smith (1983), we used the following affine gap penalties: +1 for match, 0 for mismatch, -3 to open a gap, and -0.1 to extend a gap by each extra amino acid. Informal testing revealed little sensitivity of our results to the specific values chosen for these parameters. Because alignment scores are proportional to length, each score was normalized by dividing it by the minimum of the lengths of the two sequences being aligned.

The baseline accuracy of fold-class prediction by homology modeling within our data set when protein sequences were represented by their original amino acid identities was 77.7%. However, when the sequences were transformed by the Dayhoff partition and the pairwise alignment scores were re-computed, the accuracy dropped to 56.4%. This

Table 4. The data set used for the experiments in this paper. The fold classifications are taken from Pascarella and Argos (1992), and sequences are given by their PDB names (Bernstein et al., 1977). Optional chain and sequence limits are indicated in parentheses where domains were specified.

fold	sequences
256B	2ccy(A),256b(A)
AC-PROT	1cms(1-175),1cms(176-323),4ape(2-174),4ape(175-323), 2apr(1-178),2apr(179-325),4pep(1-174),4pep(175-326)
BARREL	2taa(A),1wsy(A),1tim(A),1gox,1ypi(A),1fcb(A:100-511)
BINDING	2liv,2lbp,2gbp
CARBONIC	1ca2,2cab
CA-BIND	3cln,5cpv,3icb,4tnc,5tnc
GCR	2gcr(1-39),2gcr(40-86),2gcr(87-127),2gcr(128-173)
CYTB	1fcb(A:1-99),3b5c
CYTC	451c,1ccr,1cyc,5cyt(R),3c2c,155c(1-122)
CYT3	1cy3,2cdv
DFR	3dfr,4dfr(A);8dfr,1dfh(A)
EGLIN	1cse(I),2ci2(I)
FAD-NADH	1phh,3grs(18-161),3grs(186-294)
FCX	3fcx,1ubq
GLOBIN	4hhb(A),4hhb(B),2mhb(A),2mhb(B),1fdh(G),1mbd,1mbs, 2lhb,1eca,2lh1,1pmb(A),1mba
HLA-A2	3hla(A:1-90),3hla(A:91-182)
IGB	2fb4(L:1-109),2fb4(L:110-214),2fb4(H:1-118),2fb4(H:119-221), 2fbj(L:1-106),2fbj(L:107-213),2fbj(H:1-118),2fbj(H:119-218), 1mcp(L:1-113),1mcp(H:1-122),1rei(A),2rhe,2hfl(L:1-105), 2hfl(H:1-116),2hfl(H:117-213),1cd4,3hla(A:183-270), 3hla(B:1-99),4fab(L:1-112),3hfm(L:1-108),1mcw(W)
IL	1i1b(3:51),1i1b(50:107),1i1b(108-153)
INHIBIT	1tgs(I),3sgb(I),2ovo,1ovo(A)
LTN	3cna,2ltn(A:1-108),2ltn(A:109-181),2ltn(B:1-47)
LZM	3lzm,2lzt,2lz2,1lz1,1alc
NBD	4mdh(A),2ldb,1ldm,5ldh,1ldx,1llc,8adh,3gpd(R),1gpd(G), 1gpd1(O),1fx1,4fxn,2sbt,3adk,8atc
PLIPASE	1pp2(R),1bp2,1p2p
KINASE	1pfk(A),3pfk
PLASTO	2paz,1azu,2aza(A),7pcy
RDX	4rxn,1rdg,6rxn
REPRESSOR	1r69,2cro
RHD	1rhd(1-146),1rhd(152-293)
SBT	1cse(E),1sbt,1tex(X),2prk
S-PROT	1ton,2ptn,2trm,4cha(A),3est,1hne(E),1sgt,2sga,3sgb(E),2alp
TOX	2abx(A),1nxb,1ctx
VIRUS	4rhv(1),4rhv(2),4rhv(3),4sbv(A),2mev(1),2mev(2),2mev(3), 2tbv(A),2stv,2plv(a),2plv(2),2plv(3),1r1a(1),1r1a(2),1r1a(3)
VIRUS-PROT	3hpv,2rsp(A)
WGA	7wga(A:1-43),7wga(A:44-86),7wga(A:87-129),7wga(A:130-171), 9wga(A:1-43),9wga(A:44-86),9wga(A:87-129),9wga(A:130-171)
XIA	4xia(A),3xia

Table 5. During each iteration, our direct hill-climbing program generated and evaluated all possible perturbations of the current best partition, starting with the Dayhoff partition. The one that produced the greatest increase in accuracy was selected as the basis for the next iteration. After three iterations, no further improvements were possible. This table shows the evolution of the partition.

iteration	accuracy	partition
(Dayhoff)	56.4	{ {MILV} {FYW} {HRK} {ACDEGNPQST} }
1	72.1	{ {MILV} {FYW} {HRK} {ADEGNPQST} {C} }
2	77.1	{ {MILV} {FYW} {HRK} {ADEGNPST} {C} {Q} }
3	77.7	{ {MILV} {FYW} {HR} {ADEGNPST} {C} {QK} }

result is˜somewhat surprising in light of the intuitive relevance of Dayhoff's amino acid groupings to protein structure.

Some of the classifications of amino acids in this particular partition are questionable, however. Another partition may group amino acids together in a better way. Other nodes in this space may be constructed by swapping an amino acid from one class to another. When two partitions differ only in the class membership of a single amino acid, we define this as a *perturbation.*

To search for partitions that increase the predictive accuracy when applied as a representation, we implemented a program to hill-climb through this space (Winston, 1984), starting from the Dayhoff partition (see Table 5). Given that each of the 20 amino acids can be swapped from its initial class to one of three others, or possibly to a new unique class, there are 80 perturbations of the Dayhoff partition. Our program generated each of these variant partitions and evaluated them by their effects on the accuracy of fold-class prediction. The best change in representation was found to be the one in which cysteine was split out of the small class into its own singleton class. This revision increased accuracy from 56.4% to 72.1%. The uniqueness of cysteine was also observed by Dayhoff et al. (1972), and makes biophysical sense because it is the only amino acid that can form disulfide bridges (Richardson & Richardson, 1989).

In the next iteration of hill-climbing, our program split glutamine out of the small class into its own class, increasing the accuracy to 77.1%. This result is rather unexpected because glutamine shares properties with several other members of the small group. Specifically, glutamine contains a polar amide group like asparagine, and is roughly the same shape as glutamate. The uniqueness of glutamine might be explained by its particularly flexible *and* polar side-chain (Richardson & Richardson, 1989).

The third hill-climbing step our program took was to split lysine out of the positive class into the new class containing glutamine, which brought the accuracy back up to the baseline for using amino acid identities: 77.7%. The partial similarity between the hydrophobic methylene groups in lysine and glutamine supports the above explanation, despite their differences in size and absolute charge. After three iterations, hill-climbing halted because no perturbation of the final partition improved the accuracy.

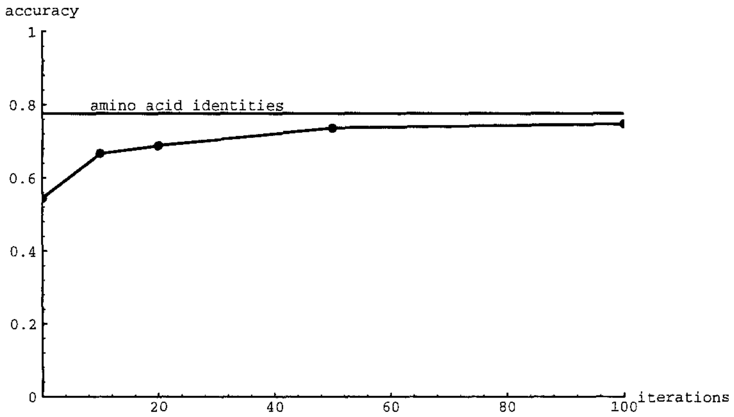

Figure 2. Accuracy of fold-class prediction for different amounts of search, based on homology modeling within our 179-protein data set. Each data point is an average over 10 runs of 10-fold cross-validation; the standard errors were too small to be displayed (less than 1%). The starting representation (0 iterations) was the Dayhoff partition. For comparison, the line marked "amino acid identities" is the baseline accuracy achieved by using the original representation of the protein sequences. A data point was computed for 200 iterations, but is not shown here because it did not increase the accuracy over the 100-iteration score.

6.1. Searching Through Partitions

Since direct hill-climbing terminated after only three iterations, we devised a related search technique, implemented in a second program, to explore more nodes in the search space. At each step, a random perturbation was generated by swapping some amino acid from its current class to another class, or to a new singleton class with 20% probability. If this perturbed partition increased accuracy at all, it was selected as the basis for the next iteration. This technique allows sub-optimal steps to be taken, which may help the search to avoid dead-end paths due to local maxima.

Because of the randomness in the algorithm, the effectiveness of this search had to be evaluated over multiple paths. Thus our second program averaged the improvement in accuracy resulting from a given number of iterations over 10 independent runs. In addition, 10-fold cross-validation was used to guard against overfitting.

Figure 2 shows the relationship between the number of iterations of search and the accuracy of fold-class prediction when the resulting partitions were used to transform example sequences into new representations. After 100 iterations, the program had improved the accuracy from 56.4% (for the Dayhoff partition) to 74.8% (on average; the standard error for each data point was around 0.5%). The improvement leveled off at this point; even with 200 iterations of search, the program did not reach a significantly higher accuracy on average (data not shown).

Although this search did not surpass the baseline accuracy (77.7% for amino acid identities) on average, the program did generate some individual partitions that were in fact better representations. The single best partition discovered was: {{MILV}{KRF}{APT}{CQN}{SW}{D}{E}{G}{H}{Y}}. On the entire data set, the accuracy of fold-class

Table 6. The 11 best partitions after 100 iterations of search.

accuracy	partition
81.0	{ {MILV} {Y} {KRF} {APT} {CQN} {SW} {D} {E} {G} {H} }
80.5	{ {MILV} {FWY} {HS} {AT} {EQ} {N} {G} {K} {CP} {R} {D} }
79.9	{ {MILV} {FWY} {HKP} {DNST} {A} {E} {GQ} {C} {R} }
79.9	{ {MILVW} {CF} {KQRY} {HNST} {P} {E} {AG} {D} }
79.9	{ {MILV} {FWY} {KQR} {ADNST} {P} {C} {EG} {H} }
79.3	{ {MILV} {NWY} {HR} {AKS} {DEQ} {G} {C} {F} {T} {P} }
79.3	{ {MILV} {FWY} {HKPR} {EST} {GN} {C} {A} {D} {Q} }
79.3	{ {MILVW} {Y} {AHKR} {DNST} {E} {G} {F} {P} {C} {Q} }
79.3	{ {MILV} {FWY} {HKR} {ADNQST} {P} {G} {E} {C} }
79.3	{ {MILV} {FY} {HKQR} {AENPST} {D} {G} {C} {W} }
79.3	{ {MIL} {FWY} {DKR} {AEST} {H} {G} {C} {N} {Q} {P} {V} }

prediction resulting from using this partition of amino acids as a representation in protein sequences during homology modeling was 81.0%.

This partition is interesting because it satisfies some generally accepted rules of protein structure, but breaks others. The hydrophobic class remains intact. Glutamine and asparagine, which are in the same class, both have amide groups. And alanine, proline, and threonine are all relatively hydrophobic. However, the aromatic class is split up and the positive class includes phenylalanine. Because this partition is so much better for homology modeling than the Dayhoff partition, the meanings of these novel relationships need to be explored.

6.2. Analyzing Search Results

At the end of 10 runs of 10-fold cross-validation, the search had produced 100 independently improved partitions. When used to re-represent protein sequences, these partitions produced fold-class prediction accuracies ranging from 72.1% to 81.0% (measured over the whole data set). Table 6 shows the 11 best partitions, all of which had accuracies greater than 79.0%. Several patterns emerge. The hydrophobic residues appear to be relatively stable as a class, whereas the small class is significantly fractured. Glutamine appears with lysine in three of the 11 partitions, which is consistent with the results of hill-climbing discussed above.

By examining the patterns of amino acid relationships in all 100 improved partitions, we were able to extract refinements of the initial domain knowledge. The first step in analyzing this data was to compute the frequencies with which each pair of residues was found in the same class. As expected, hydrophobic residues were found to be in the same class most often; each of the six pairs within this set of four residues was found to occur in the same class in at least 88 of the 100 partitions. Similarly, the positive and aromatic groups were clearly defined by the pairwise frequencies of members of these groups to be found together in the 100 partitions, while other amino acids tended not to associate with them.

The class of small residues was least well-defined. Interestingly, six of the original ten residues formed a fairly clear subgroup consisting of alanine, aspartate, serine, glutamate, asparagine, and threonine. Thirteen of the fifteen pairs within this subgroup were found to be associated in the partitions more frequently (at least 40% of the time) than any pair formed by a group member and a non-group member. Thus we take this subset of amino acids as the core definition of the small group: {ADENST}.

Our second step in analyzing this data was to compute the frequency with which specific amino acids associated with the original amino acid groups. Since the classes in the Dayhoff partition (see Table 3) might have been altered due to perturbations during search, we identified representatives of these groups of residues by classes that contained a majority of them. For example, a class such as {KRF} is a surrogate positive group because it contains two of the three original positive residues from {HRK}. Any class that contained four or more of the six core small residues was considered to represent the small class.

Given this method for identifying representatives of the original Dayhoff groups in arbitrary partitions, we collected the frequencies with which each amino acid was found in each of these groups among the 100 perturbed partitions (see Figure 3). Occasionally an amino acid occurred in a group that did not appear to represent any of the original classes. A significant number of these cases were accounted for by amino acids in their own singleton classes. Thus the histograms also show scores for both unique classes and unanalyzable cases.

These histograms can be interpreted as preferences for the amino acids to express certain properties, and can thus be considered as refinements of the initial knowledge. For example, the strong hydrophobicity of methionine, leucine, valine, and isoleucine is apparent. Although not all researchers include methionine in the hydrophobic group because it has a polar sulfer atom (Taylor, 1986), these results show that, for the purposes of homology modeling, methionine definitely belongs in this group. Similarly, some researchers include histidine in the aromatic group (Taylor, 1986). However, this data shows no preference for histidine to associate with the aromatic group.

Subtle patterns emerge in comparisons of the histograms between similar amino acids. For example, the greater hydrophobicity of threonine over serine is evident. And of the aromatic residues, tryptophan is most hydrophobic, while tyrosine is least hydrophobic.

Perhaps the most interesting aspect of these histograms is the propensity of cysteine, glycine, proline, and glutamine to split out of the class of small residues into their own singleton classes. There are clear reasons why three of these four residues should be considered unique. Cysteine can form disulfide bonds, glycine allows extra flexibility in the protein backbone, and proline restricts the backbone torsion angles (Richardson & Richardson, 1989). However, glutamine seems to share many properties with other residues in the small class, and, as in the hill-climbing experiment above, its uniqueness is unexpected.

Overall, our results generally follow the patterns of amino acid relationships expressed in the original Dayhoff partition, except for the refinement of the small class of residues. The fact that four unique residues were split out of this group de-emphasizes the significance of smallness as a property for homology modeling. The most general partition

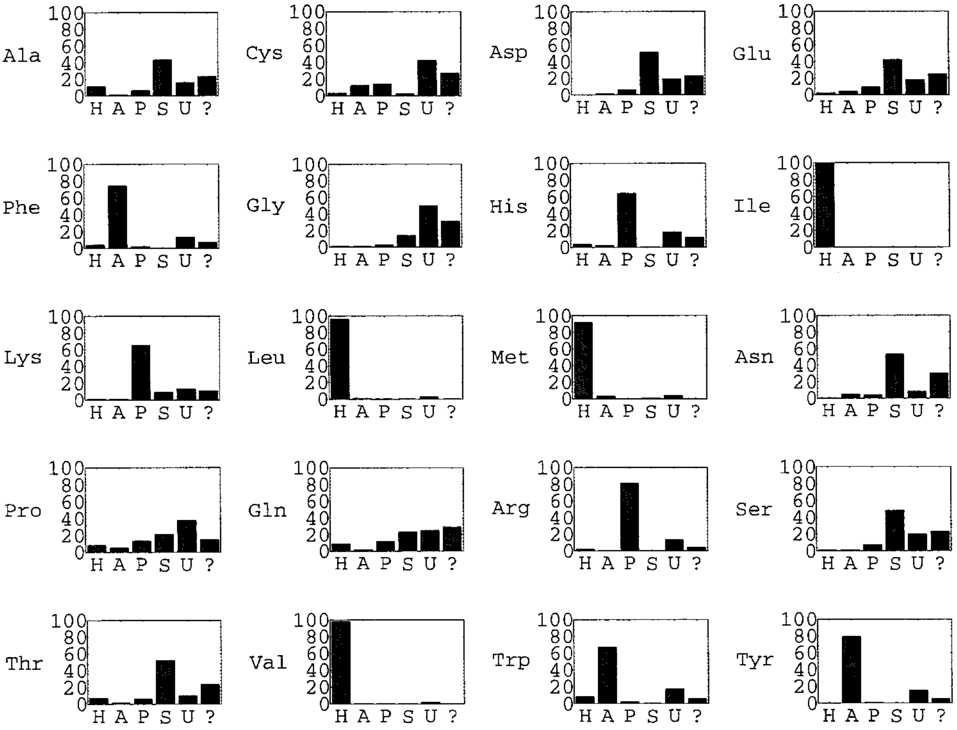

Figure 3. Histograms of the group preferences for each of the amino acids, summarizing the frequencies of association between each amino acid and representatives of the Dayhoff groups among the 100 perturbed partitions. The column letters represent amino acid groups: H=hydrophobic, A=aromatic, P=positive, S=small, U=unique, and ?=unanalyzable.

suggested by these analyses is {{MILV}{HRK}{FWY}{ADENST}{C}{G}{P}{Q}}, and its accuracy is computed to be 77.1%, which is just below the baseline accuracy for using amino acid identities (77.7%).

Thus our method for using domain knowledge to construct and search a space of alternative representations has not only improved the accuracy of fold-class prediction (from 77.7% to 81.0% on our data set), but analysis of the search results has suggested an important revision of the domain knowledge. Future work will include searching for refinements of other amino acid groups, further exploration of the best partitions (e.g. those in Table 6), and an analysis of the sensitivity of these results to the particular data set we used for these experiments.

6.3. Context-Dependence of Amino Acid Relationships

While partitions can capture many variations of amino acid groupings, they are fundamentally limited in expressive power. Because groups of amino acids cannot overlap, a commitment must be made as to the group membership for each amino acid. Thus this formalism cannot exploit the multiple dimensions of similarity among the amino acids which arise from their variety of properties. This is another possible explanation for why the search procedure introduced in Section 6.2 did not surpass the baseline accuracy on average (see Figure 2).

However, domain knowledge suggests that the various roles an amino acid can play might be context-dependent. Because an amino acid usually participates in determining the local structure by interacting with nearby residues, the property that is playing a structural role at a given site might be related to the properties of neighboring amino acids (Overington et al., 1992).

We propose that, by extending the representation of amino acid sequences to take sequential context into account, the expressive power can be increased to capture finer relationships among the amino acids. We can construct new sequence elements by taking the cross-product of the central amino acid at each site with the properties of the amino acids surrounding it. This transforms each amino acid symbol in the sequence into a specialized case that depends on its context.

An example of this method of re-representing protein sequences is to include the hydropathicity of the two closest neighbors for each amino acid. Thus an alanine in the context of a particular site in a particular sequence, can really be one of four distinct cases: alanine surrounded by two hydrophobic residues, alanine surrounded by two hydrophilic residues, alanine with a hydrophobic N-terminal neighbor and a hydrophilic C-terminal neighbor, and alanine with a hydrophilic N-terminal neighbor and a hydrophobic C-terminal neighbor.

For brevity, we use an extended set of sequence symbols, such as ⟨i ala i⟩, ⟨o ala o⟩, ⟨o ala i⟩, and ⟨i ala o⟩, where 'i' stands for hydrophilic and 'o' stands for hydrophobic. These symbols replace single amino acids in a protein sequence. For example, the sequence ...His Lys Ala Leu Arg... would become ...⟨? his i⟩ ⟨i lys o⟩ ⟨i ala o⟩ ⟨o leu i⟩ ⟨o arg ?⟩.... We take the following amino acids to be hydrophobic for our experiments:

{MILVAFGP}, and we assume that terminal residues are adjacent to one extra hydrophilic residue.

Because each sequence initially gets transformed from an alphabet of 20 amino acids to an alphabet of 80 context-based symbols, the elements of the partitions must be multiplied into this many cases as well. For example, if one of the classes in a partition is {MILV}, the corresponding class in the extended partition would be {⟨o met o⟩ ⟨i met i⟩ ⟨o met i⟩ ⟨i met o⟩ ⟨o ile o⟩ ⟨i ile i⟩ ⟨o ile i⟩ ⟨i ile o⟩ ⟨o leu o⟩ ⟨i leu i⟩ ⟨o leu i⟩ ⟨i leu o⟩ ⟨o val o⟩ ⟨i val i⟩ ⟨o val i⟩ ⟨i val o⟩}.

The advantage of this representational extension is that amino acids in different contexts can be swapped independently into different classes. So if threonine acts like valine in a hydrophobic context, but otherwise tends to exchange with serine, then ⟨o thr o⟩ can be placed in the class with the symbols for valine, while the other three cases can be grouped with the symbols for serine. Thus the expressive power has been increased to capture more subtle relationships among the amino acids.

To explore the utility of our proposed extension to the representation of amino acid sequences, we first formed the context-based extension of the revised Dayhoff partition (with the four unique residues split out of the small class; accuracy still 77.1%). Then we wrote a program to construct a restricted set of perturbations of this partition by allowing any combination of the context-dependent cases for a particular residue to distribute among a set of alternative classes.

We used this program to test the context-dependence of the role of glycine, which the histograms in Figure 3 suggest is both small and unique. The program generated 16 partitions based on whether each of the four extended symbols for glycine was placed in the small class or the unique class for glycine. By evaluating each of these partitions, our program found the best representation for homology modeling to be when only ⟨i gly o⟩ was grouped with the small residues. This increased the accuracy by more than one percentage point to 78.2%.

Similarly, we used this search procedure to explore the context-dependence of the relationship between glutamine and lysine. Our direct hill-climbing procedure (see Table 5) placed glutamine in its own class, and then swapped lysine into this class, suggesting that they often play a common role in protein structures. We used our context-based search program to construct the 16 variant partitions in which the symbols for glutamine were distributed independently between the positive class and the unique class for glutamine. Our program found that the best grouping was to place ⟨i gln o⟩ and ⟨i gln i⟩ with the symbols for lysine in the positive class, increasing accuracy to 79.3% (an improvement of more than 2% over the accuracy for the revised Dayhoff partition). This result suggests that glutamine behaves like a positive residue when its N-terminal neighbor is hydrophilic.

These initial results are promising because they demonstrate that significant improvements in accuracy can be achieved with the more expressive representation. Because contextual information is incorporated directly into the representation of sites in a protein, this extended representation can capture multiple dimensions of similarity among the amino acids.

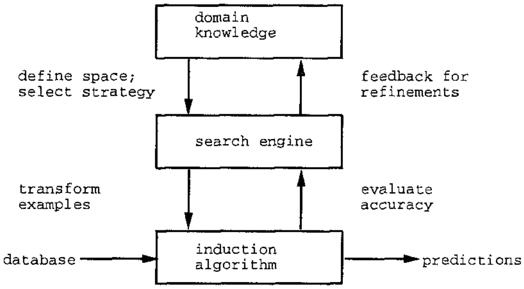

Figure 4. An abstract framework for searching for representations to improve machine learning performance. At the lowest level is a performance element which is doing induction in the domain. At the highest level, knowledge suggests a space of alternative representations. At the intermediate level, search in conducted through this space. Representations are evaluated by transforming a database of training examples and estimating the accuracy of prediction. Finally, the results of the search can be used as feedback to refine the high-level domain knowledge.

However, the best context-based partition that combined the two above results, placing ⟨i gly o⟩ in the small class and ⟨i gln o⟩ and ⟨i gln i⟩ in the positive class, only produced an accuracy of 77.1%. The lack of additivity suggests that there are more complex interactions in this extended space of representations, which might necessitate a more sophisticated search strategy. Future work will include the of use a genetic algorithm to search several paths in parallel (Holland, 1975).

7. Discussion

In this paper, we have presented a specific method for improving protein fold-class prediction by re-representing protein sequences in terms of amino acid properties. A generalization of this technique can be applied to improve machine learning in other difficult, real-world domains. In this section, we propose an abstract framework for using domain knowledge to search for representations that improve learning performance.

Our framework has several components, depicted in Figure 4. At the lowest level is standard induction. We assume the domain has been sufficiently formalized to identify objects to be classified (Michalski, 1983). The language for describing training examples constitutes the initial representation of objects. Furthermore, we assume there is an established technique for making predictions, whether it is a domain-specific method or some standard learning algorithm that achieves the best accuracy in the domain to date.

The highest level of our framework consists of domain knowledge. The knowledge does not have to be encoded in any particular form, such as a Horn-clause theory. However, the knowledge must be applicable to the representation. Our framework requires domain knowledge to suggest ways in which the representation of examples can be extended. The closure of all potentially relevant extensions to the representation defines a space of alternatives to the initial language for describing examples.

The middle level of our framework mediates between the domain knowledge and induction. This level consists of a search engine for exploring the space of alternative representations suggested by the domain knowledge. Domain knowledge is also needed to select a search strategy. For example, the relative uncertainties of pieces of knowledge used to construct the search space can suggest which regions are most likely to contain an improved representation, and hence should be searched first.

After the search strategy selects a particular representation as a node in the search space, it is evaluated using the inductive component. The representation is applied to a set of training examples, and the accuracy of prediction is estimated by cross-validation using the re-represented database.

This search framework abstracts our approach to improving protein fold-class prediction. At the lowest level, homology modeling represents the inductive algorithm. At the highest level, domain knowledge suggests that partitions of amino acids might capture relationships among the amino acids that are relevant to protein structure prediction. The set of partitions defines a space of alternative representations to be searched, and several versions of hill-climbing were tried in our experiments. When a partition was constructed during search, it was applied to re-represent a set of example sequences, and the partition was evaluated according to the resulting accuracy of homology modeling.

This framework can also be applied to other difficult, real-world domains, provided that knowledge be can expressed in a form that is relevant to representation. Background knowledge often suggests patterns to look for, intermediate concepts that might be useful, or potential interactions among features. Despite the typical uncertainty in this knowledge, it can often be expressed via an extension of the representation. For example, membership in an intermediate concept can be computed for each example, and this value can be appended to their descriptions. Similarly, if two features are suspected to interact in a particular way, then a third feature that combines them can be computed and used to extend the description for each example (Matheus, 1989).

The importance of this framework is that it reveals several systematic roles for knowledge to play in machine learning. Many real-world domains, such as weather prediction (Packard, 1989), speech production (Sejnowski & Rosenberg, 1987), and financial valuation (Ragavan et al., 1993), are said to be difficult because the best known algorithms can only achieve a limited accuracy. In some cases, even small improvements in accuracy over competitive methods can provide significant advantages. For this reason, it is crucial to exploit any background knowledge, which often exists in real-world domains.

However, there is no comprehensive theory of the interaction between knowledge and learning. Presumably, knowledge could be used to construct an induction algorithm with a domain-specific bias (Mitchell, 1980), although this is hard to do in practice. One approach to incorporating knowledge into learning has been to use an explicit domain theory to guide the generalization of examples by explanation (DeJong & Mooney, 1986).

In our framework, knowledge is first used to define a space of alternative representations, and then to select a strategy for searching this space. The representation of examples in a domain has a strong effect on the performance of a learning algorithm (Rendell & Seshu, 1990). Representation interacts with the algorithmic bias to determine which generalizations are drawn from a set of training examples (Mitchell, 1980). A pervasive algorithmic bias, such as the propensity of similarity-based learning (SBL) algorithms to focus on contiguous regions of instance space (Rendell & Ragavan, 1993), can cause the representation in a domain to appear to be "low-level" in general. Our framework extends the utility of existing learning algorithms in difficult domains by adapting the representation to the inductive bias.

Several previous methods have used change of representation to improve machine learning performance. Michalski (1983) introduced *constructive induction* operators to change the representation of logical descriptions in INDUCE. Utgoff (1986) implemented another method in STABB, calling this general approach *dynamic bias.* And *feature construction* has always be seen as a way of extending the representation in a feature-vector-based domain (Matheus, 1989). Our framework unifies these approaches in terms of their search for alternative representations to improve the performance of an underlying inductive algorithm.

An important aspect of our approach is the use of a database of training examples to evaluate and compare representations. Chrisman (1989) and Cohen (1990) proposed formal frameworks for improving learning by change of representation, but they were both based on an exact-learning model. Our framework exploits incremental increases in accuracy to guide the search process. In fact, our approach can be thought of as "tuning" the representation to the bias of the preferred algorithm, which is equivalent to *bias optimization* (Tcheng et al., 1989). This use of examples is similar to the performance evaluations of several widely used amino acid similarity matrices by Henikoff and Henikoff (1993), although they did not search beyond the initial set of matrices being compared.

One advantage of our search framework is that it also suggests a reverse interaction between knowledge and learning. While knowledge is input to the search engine to define the space and strategy, refinements of the knowledge might also be extracted from the search results. In the protein fold-class domain, we showed how a set of improved partitions can be summarized to detect patterns of change from the initial representation, and this analysis was used to refine the Dayhoff partition.

In general, the search results can provide feedback on the knowledge by evaluating its utility for learning. If a particular region in the space of alternative representations is recommended for search by some piece of knowledge, but a brief exploration of those representations does not produce any improvements in accuracy, then the certainty in that piece of knowledge might be decreased. This feedback can be used by an expert to revise the domain knowledge, specifically guided by the constraints of the prediction task itself. This approach based on the connection between knowledge and representation contrasts with automated systems for theory revision, such as KBANN (Towell et al., 1990), which typically assume that the knowledge can be used to explain the classification of examples directly.

8. Conclusion

We have presented a method for improving protein fold-class prediction based on re-representing protein sequences in terms of amino acid properties. Experimental results showed that, not only can this method increase the accuracy of predicting protein structures, but it can also be used to refine biophysical knowledge of the various roles amino acids can play in determining protein structure. Specifically, on one data set, accuracy was improved from 77.7% to 81.0%, and analysis suggested that cysteine, glycine, proline, and glutamine should be split out of the group of small amino acids.

We then generalized our domain-specific method to an abstract framework in which machine learning performance is improved by searching for better representations. Representations are evaluated for comparison by transforming a set of training examples, applying a selected induction algorithm, and estimating predictive accuracy. The framework provides systematic roles for using domain knowledge to improve learning in difficult, real-world domains: knowledge is used to define a space of alternative representations and to select a strategy for searching that space. This framework also suggests that refinements of the knowledge may be extracted by summarizing the results of the search.

Acknowledgments

This research was supported in part by grants NSF ASC-89-02829 (SS) and NSF IRI-92-04473 (LAR), and a Beckman Institute graduate research assistantship (TRI). We would like to thank the National Center for Supercomputing Applications for providing computational resources. We would also like to thank the anonymous *Machine Learning* reviewers.

References

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning, 6*, 37–66.

Baldwin, R. L. (1989). How does protein folding get started? *Theoretical Issues in Biological Sciences, 14*, 291–294.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology, 112*, 535–542.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature, 326*, 347–352.

Chothia, C. (1988). The fourteenth barrel rolls out. *Nature, 333*, 598–599.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature, 357*, 543–544.

Chothia, C. and Lesk, A. M. (1986). The relation between divergence of sequence and structure in proteins. *The EMBO Journal, 5*, 823–826.

Chou, P. Y. and Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry, 13*, 222–244.

Chrisman, L. (1989). Evaluating bias during PAC-learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 469–471. Palo Alto, CA: Morgan Kaufmann Publishers.

Cohen, W. W. (1990). An analysis of representation shift in concept learning. In *Machine Learning: Proceedings of the Seventh International Conference*, pages 104–112. Palo Alto, CA: Morgan Kaufmann Publishers.

Dayhoff, M., Eck, R., and Park, C. (1972). A model of evolutionary change in proteins. In Dayhoff, M., editor, *Atlas of Protein Sequence and Structure*, volume 5. Silver Spring, MD: National Biomedical Research Foundation.

DeJong, G. F. and Mooney, R. J. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145–176.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry, 29*, 7133–7155.

Doolittle, R. F. (1981). Similar amino acid sequences: Chance or common ancestry? *Science, 214*, 149–159.

Doolittle, R. F. (1986). *Of Urfs and Orfs: A Primer on How to Analyze Devised Amino Acid Sequences*. Oxford University Press: Oxford.

Finkelstein, A. V. and Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns. *Progress in Biophysics and Molecular Biology, 50*, 171–190.

Fitch, W. M. and Smith, T. F. (1983). Optimal sequence alignments. *Proceedings of the National Academy oj Sciences, USA, 80*, 1382–1386.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology, 162*, 705–708.

Gribskov, M., Homyak, M., Edenfield, J., and Eisenberg, D. (1988). Profile scanning for three-dimensional structural patterns in protein sequences. *CABIOS, 4*, 61–66.

Henikoff, S. and Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins, 17*, 49–61.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press: Ann Arbor, MI.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature, 358*, 86–89.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry, 4*, 23–54.

King, R. and Sternberg, M. (1990). Machine learning approach for the prediction of protein secondary structure. *Journal of Molecular Biology, 216*, 441–457.

Lathrop, R. H., Webster, T. A., and Smith, T. F. (1987). Pattern-directed and hierarchical abstraction in protein structure recognition. *Communications of the Association for Computing Machinery, 330*, 909.

Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science, 227*, 1435–1441.

Matheus, C. (1989). *Feature Construction: An Analytic Framework and an Application to Decsion Trees*. PhD thesis, University of Illinois, Department of Computer Science.

McCammon, J. and Harvey, S. (1987). *Dynamics of Proteins and Nucleic Acids*. New York: Cambridge University Press.

McLachlan, A. D. (1972). Gene duplication in carp muscle calcium-binding protein. *Nature New Biology, 240*, 83–85.

Michalski, R. (1983). A theory and methodology of inductive learning. *Artifical Intelligence, 20*, 111–161.

Mitchell, T. (1980). *The Need for Biases in Learning Generalizations*. Technical Report CBM-TR-117, Rutgers: New Brunswick, NJ.

Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *CABIOS,4* , 11–17.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*, 443–453.

Neidhart, D. J., Kenyon, G. L., Gerlt, J. A., and Petsko, G. A. (1990). Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature, 347*, 692–694.

Nell, L. J., McCammon, J. A., and Subramaniam, S. (1992). Anti-insulin antibody. Structure and conformation I. Molecular modeling and mechanics. *Biopolymers, 32*, 11–21.

Overington, J., Donnelly, D., Johnson, J. S., Sali, A., and Blundell, T. (1992). Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Science, 1*, 216–226.

Packard, N. H. (1989). Genetic learning algorithm for the analysis of complex data. Center for Complex Systems Research Report CCSR-89-10, University of Illinois: Urbana, IL.

Pascarella, S. and Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Engineering, 5*, 121–137.

Qian, N. and Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology, 202*, 865–884.

Ragavan, H., Rendell, L., Shaw, M., and Tessmer, A. (1993). Complex concept acquisition through directed search and feature caching. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 946–951.

Rendell, L. and Ragavan, H. (1993). Improving the design of induction methods by analyzing algorithm functionality and data-based complexity. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 952–958.

Rendell, L. and Seshu, R. (1990). Learning hard concepts through constructive induction: Framework and rationale. *Computational Intelligence, 6*, 247–270.

Richards, F. (1992). Folded and unfolded proteins: An introduction. In Creighton, T., editor, *Protein Folding*, pages 1–58. Freeman: New York.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry, 34*, 167–336.

Richardson, J. S. and Richardson, D. C. (1989). Principles and patterns of protein conformation. In Fasman, G. D., editor, *Prediction of Protein Structure and the Principles of Protein Conformation*, pages 1–98. New York: Plenum Press.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins, 9*, 56–68.

Schulz, G. E. and Schirmer, R. H. (1979). *Principles of Protein Structure.* Springer-Verlag: New York.

Schwartz, R. M. and Dayhoff, M. O. (1978). Matrices for detecting distant relationships. In Dayhoff, M., editor, *Atlas of Protein Sequence and Structure*, volume 5, supplement 3. Silver Spring, MD: National Biomedical Research Foundation.

Sejnowski, T. J. and Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English texts. *Complex Systems, 1*, 145–168.

Smith, R. F. and Smith, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Biochemistry, 87*, 118–122.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*, 195–197.

Stryer, L. (1988). *Biochemistry.* W. H. Freeman and Company: New York.

Subramaniam, S., Tcheng, D., Hu, K., Ragavan, H., and Rendell, L. (1992). Knowledge engineering for protein structure and motifs: Design of a prototype system. In *Proceedings of the Fourth International Conference of Software Engineering and Knowledge Engineering*, pages 420–433. IEEE Computer Society: Washington, DC.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *Journal of Molecular Biology, 188*, 233–258.

Tcheng, D. K., Lambert, B. L., Lu, S. C. Y., and Rendell, L. A. (1989). Building robust learning systems by combining induction and optimization. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 806–812.

Towell, G., Shavlik, J., and Noordewier, M. (1990). Refinement of approximate domain theories by knowledge-based neural networks. In *Proc. Eighth Natl. Conf. on Artificial Intelligence*, pages 861–866.

Utgoff, P. (1986). Shift of bias for inductive concept learning. In Michalski, R., Carbonell, J., and Mitchell, T., editors, *Machine Learning: An Artificial Intelligence Approach, II*, pages 107–148. San Mateo, CA: Morgan Kaufmann Publishers.

Watson, J. D. (1990). The human genome project: Past, present, and future. *Science, 248*, 44–49.

White, F. H. (1961). Regneration of native secondary and tertiary structures by air oxidation of reduced ribonuclease. *Jounral of Biological Chemistry, 236*, 1353–1360.

Winston, P. (1984). *Artifical Intelligence.* Reading, MA: Addison-Wesley.

Received October 7, 1993
Accepted November 22, 1994
Final Manuscript January 16, 1995