Machine Learning, 21, 125–150 (1995)

Machine Discovery of Protein Motifs

DARRELL CONKLIN * conklin@zgi.com
ZymoGenetics Inc., 1201 Eastlake Avenue East, Seattle, WA 98102

Editors: Jude Shavlik, Lawrence Hunter, and David Searls

Abstract. The investigation of relations between protein tertiary structure and amino acid sequence is a topic of tremendous importance in molecular biology. The automated discovery of recurrent patterns of structure and sequence is an essential part of this investigation. These patterns, known as *protein motifs*, are abstractions of fragments drawn from proteins of known sequence and tertiary structure. This paper has two objectives. The first is to introduce and define protein motifs, and provide a survey of previous research on protein motif discovery. The second is to present and apply a novel approach to protein motif representation and discovery, which is based on a *spatial description logic* and the symbolic machine learning paradigm of structured concept formation. A large database of protein fragments is processed using this approach, and several interesting and significant protein motifs are discovered.

Keywords: protein tertiary structure, machine discovery, relational learning, knowledge representation, description logics, information retrieval, knowledge discovery in databases

1. Introduction and definitions

A task of growing importance in the management of biochemical and crystallographic data is the ability to perform generalization and abstraction over large sets of related physical observations. There exist recurrent patterns and rules of structural biochemistry hidden in the Protein Data Bank (Bernstein et al., 1977); machine discovery techniques can help to uncover these patterns and rules. Generalized patterns can facilitate efficient information retrieval and data incorporation, providing a conceptual framework to which new structural data can be related. They can also be used for prediction of molecular conformation from topological structure, since they often represent common 3D (three-dimensional) structural features. The analogous investigation of relations between protein structure and amino acid sequence is a topic of tremendous importance in molecular biology. It is necessary to understand protein 3D structure before protein function and activity can be understood at the molecular level.

Proteins are macromolecules comprising chains of structural building blocks known as *amino acids*. Amino acids have a *backbone* and a *side chain*, and are classified into 20 groups based on the topological structure of their side chain. Each amino acid is typically denoted by a single letter name. Proteins can be described at various levels of abstraction. The *primary structure* of a protein is given by its linear (1D) sequence of amino acids. The primary structure dictates the 3D structure of the protein in solution, although the rules governing this determination have not yet been discovered. The *secondary structure* of a protein is given by a sequence of structural identifiers, such as H (alpha–helix), E

* This work was performed while the author was at Queen's University, Kingstson, Ontario, Canada.

```
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN
-EE-SSHHHHHHHHHHHTTT--HHHHHHHHS-EE-SSS---GGG--
```

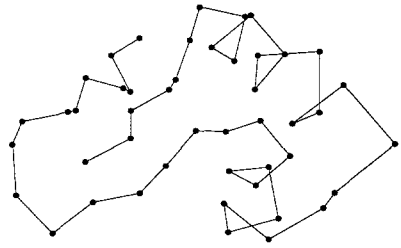


Figure 1. Representations of structure (primary, secondary, tertiary) for a small protein (Crambin: PDB code 1CRN). Top: primary structure given by a sequence of 46 amino acid identifiers. The primary sequence is followed by a sequence of 46 secondary-structure identifiers (DSSP format: a "-" indicates unassigned structure for the corresponding residue). Bottom: a 2D visual projection of the protein's tertiary structure (virtual backbone representation).

(strand), or T (turn). Equivalently, a sequence of n elements is often described by n such structural identifiers: the format produced by the DSSP secondary-structure assignment program (Kabsch & Sander, 1983). Protein *tertiary structure* is described by the positions of all atoms of the macromolecule in 3D space (see Figure 1).

In addition to the levels of secondary and tertiary structure, there are a variety of abstract representation forms for protein 3D structure (Schulz & Schirmer 1979). One such representation, encountered frequently throughout this paper, is the *protein virtual backbone*. For a chain of length n, this comprises n representative points, one for each residue. The chosen representative point is often the position of the Cα atom (one of the amino acid backbone atoms) of the residue. In a protein virtual backbone, two points are *contiguous* or *virtually bonded* if their residues are adjacent in the amino acid sequence. The *virtual bond angle* (VBA), defined between three contiguous points (a, b, c), is the angle between the vectors (b, a) and (b, c). The *virtual bond dihedral angle* (VBDA), defined between four contiguous points (a, b, c, d), is the angle between the planes (a, b, c) and (b, c, d).

A *protein fragment* is an observed pattern of amino acid residues, for example, a region of (1D) primary structure, or of (3D) tertiary structure. A *protein motif* is an abstraction of one or more fragments. Protein motifs can be classified into four categories. *Sequence* motifs are linear strings of residue identifiers with an implicit topological ordering. *Sequence–structure* motifs are sequence motifs with predefined secondary-structure identifiers attached to one or more residues in the motif. The sequence is assumed to be predictive of the associated structure. *Structure* motifs are 3D structural objects, described by positions of residue objects in 3D Euclidean space. Structure motifs are free of sequence information, although most research enforces contiguity of structure motif components. Finally, *structure–sequence* motifs are combined 1D-3D structures that associate sequence information with a structure motif. Structure–

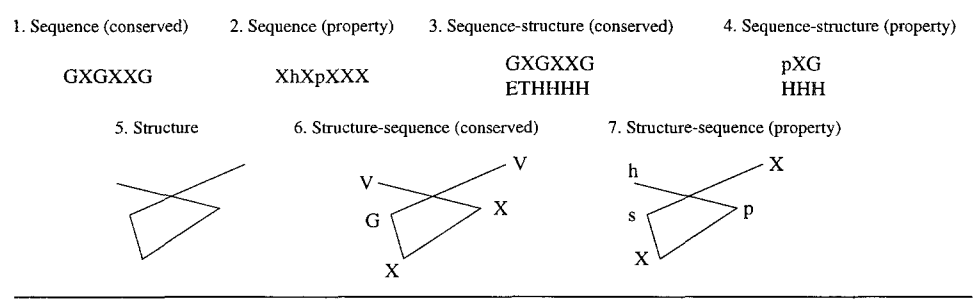


Figure 2. Various types of protein motifs. Legend; X: any residue, G: glycine, V: valine, H: helix, E: β-strand, T: turn, p:polar, s: small, h: hydrophobic.

sequence motifs need not indicate a fixed direction of implication between structure and sequence. Figure 2 illustrates these four types of protein motifs, along with some further subclassifications which are elaborated upon later in the paper. The first three motif types are discussed by Thornton and Gardner (1989); the structure–sequence motif will be presented in this paper.

Machine discovery of protein motifs of various types is currently an area of intense interest in molecular biology. One of the objectives of this paper is to bring this important application domain to the attention of the machine learning community. Section 2 of this paper will survey previous research on protein motif discovery according to the above categorization of motifs. The second aim of this paper is to present and apply a new approach to protein motif discovery, which is based on knowledge representation ideas of description logics and machine learning principles of structured concept formation. Section 3 of this paper presents this new approach. The description logic facilitates reasoning about subsumption of motifs — an ability not found in other protein motif discovery systems — while the concept formation procedure creates a concept taxonomy of generalized motifs. Section 4 presents some promising results obtained on a large database of protein fragments.

2. Discovery of protein motifs

The field of empirical machine discovery encompasses the theories and autonomous processes involved in finding novel regularities, concepts, or dependencies in data. It is convenient to identify a protein motif with a *concept*, having a formal intensional description in addition to an extensional meaning. In order to apply concept discovery techniques, a rigorous mathematical semantics for a motif is necessary to determine whether an observation lies within that concept's extension.

There has been a considerable amount of research on machine discovery of protein motifs. Most, but not all, of this work relies on some form of numerical clustering: fragments are described by a set of numeric features, a distance metric on these descriptions

is defined, and motifs emerge as cluster centroids during the clustering process. The technique presented in this paper uses, by contrast, a conceptual clustering technique; fragments are structured objects, a measure of relational or structural similarity is defined, and motifs are most specific generalizations of fragments.

This section will describe and compare a handful of methods in terms of their representation theory, the type of motif under consideration, the semantics given to motifs, the motif evaluation mechanism employed, and the pragmatics of discovered motifs. A more extensive survey appears in my Ph.D. dissertation (Conklin, 1995). Though the results presented in Section 4 will be concerned mainly with structure–related motifs, it is informative to review sequence motif work as a backdrop for discussing the more general structure–sequence motif.

2.1. Sequence and sequence–structure motifs

Protein sequence motifs are the most commonly encountered motif type in the molecular biology literature. It is generally assumed that similarities in protein sequence are indicative of structural and functional similarity. Thus the discovery of sequence motifs from structurally similar proteins or protein fragments is an important method for uncovering relationships between protein structure and sequence. Protein sequence motifs can facilitate the incremental acquisition and indexing of sequence data into knowledge bases organized according to sequence similarity (Taylor, 1986).

Sequence motifs can be discovered from a maximal alignment of one or more protein sequences, followed by the abstraction of residues at aligned positions. There is an extensive literature on the comparison of sequence motifs: see Lathrop et al. (1993) for a good survey of this work. Conserved residues are those identical at corresponding alignment positions. It is uncommon to find long connected sequences of conserved residues in non–homologous proteins (Sternberg & Islam, 1987), hence the need arises to construct histograms of residue distribution at alignment positions to produce a consensus sequence motif. Taylor (1986) uses a more general abstraction scheme; a domain theory is used to classify amino acids into nondisjoint groups based on physicochemical properties such as hydrophobicity and polarity which are expected to have an influence on protein folding. Following Rooman and Wodak (1991), we shall refer to sequence motifs containing property identifiers as *property* motifs (see Figure 2).

Much of the work on protein secondary-structure prediction is based on the *a priori* definition of sequence motifs that are predictive of a certain type of secondary-structure identifier. These sequence–structure motifs (referred to by Thornton and Gardner as "structure-related sequence motifs") manifest an inherent directionality of implication from sequence to structure. The work of Rooman, Wodak, and colleagues has been influential in establishing important sequence–structure predictivity results. Rooman and Wodak (1988) associate with each amino acid in a motif a standard secondary-structure identifier (e.g., Figure 2, motif 3). A discovery procedure develops short sequence regular expressions with structure associations. They demonstrated that, while not enough associations are derived to predict a complete protein structure, a number of reliable and predictive motifs do exist. In a similar study, Rooman et al. (1990b) replace the standard

secondary-structure identifiers with elements produced by structure motif discovery (see next section).

Rooman et al. (1989) take a number of sequence motifs, reported in the literature as characterizing local secondary structure, and subject them to a rigorous validation against a database of 75 proteins. Out of 12 sequence motifs, only one is found to be predictive of secondary structure. The hypothesized reason for this poor performance is the over-generality of the motifs tested. Rooman et al. present experimental evidence indicating that there is a direct relationship between motif specificity and predictivity.

2.2. Structure motifs

The accurate prediction of protein tertiary structure from amino acid sequence, while theoretically possible, has remained one of the great open problems in molecular biology. Though the prediction problem is typically reduced to a more manageable one — the prediction of secondary structure from sequence followed by the packing of secondary structures into 3D — recently there has been increasing interest in the automated discovery and use of structural elements less coarse than the standard secondary structures. There are three main reasons for this development. First, there is a wide discrepancy between different methods for secondary-structure assignment from tertiary structure (Zhang et al., 1993). A prediction system relying on one assignment method for training and evaluation is modelling to some extent its particular characteristics (Colloc'h et al., 1993). Second, unidentifiable folding patterns are usually classified as "random coil," even though these regions are neither random nor undefinable (Prestrelski et al., 1992). Finally, the packing of secondary-structure elements is itself a non–trivial task and is dependent on accurate secondary-structure predictions. By contrast, structure motifs are building blocks that can be used to precisely describe the tertiary structure of a new protein (Jones & Thirup, 1986).

Structure motifs are typically discovered by numerical clustering. These procedures require a fixed set of numeric parameters to describe observations, and a distance metric over these multidimensional vectors. In all structure motif discovery research surveyed in this section, the initial numeric parameters are simply the list of 3D coordinates of fragment components. The "visual" similarity of two fragments is thus highly dependent on the coordinate frame of the protein(s) from which the fragments were extracted. This problem can be circumvented in three ways: 1) by standardizing fragment descriptions before comparison; 2) by using numeric features invariant with respect to Euclidean transformations (rotations and translations), or 3) by first performing an optimal alignment of the fragments in 3D space before their comparison. There are drawbacks with each approach; the first implicitly forces a fragment alignment that is possibly suboptimal, the second is sensitive to the chosen features, and the third complicates the semantics of a motif. Further weaknesses with these approaches are discussed at the end of this section.

Hunter and States (1991) standardize fragments by placing their center of mass at the origin, orienting the x, y, and z axes relative to the principal axes of the moment of inertia tensor of the fragment. The authors note various problems with this standardization method; foremost is the fact that small changes in fragment data can flip the choice of an

axis, confounding the motif/fragment relationship. Bayesian classification techniques are used to produce clusterings which are evaluated according to Bayes' formula with a prior distribution favoring fewer classes; given two clusterings, each with the same number of classes, the method will prefer the clustering which has a tighter fit to the data. Motifs are represented by a probability distribution over Cartesian coordinates for the backbone atoms of each residue. Thus motifs are probabilistic concepts; fragments fall into a class with a certain probability. In contrast to other approaches reviewed below, amino acids are represented by *line* rather than *point* data, since all backbone atoms are used.

Rooman et al. (1990a) use an invariant representation scheme. The similarity between two fragments is computed by a Cα distance root-mean-square (DRMS) metric[1] (Rooman et al., 1990a). Similar to the work of Prestrelski et al. (1992) and Zhang et al. (1993), fixed-length contiguous fragments from a protein virtual backbone are used. A fragment is an instance of a class by virtue of being within a DRMS distance threshold from the prototypical motif of that class. Similar to the work of Hunter and States, a structure motif is a 3D coordinate description, and can be depicted and perhaps evaluated by a chemist using molecular visualization software. A fragment clustering yields a dendogram of cluster merges; this structure must be pruned to produce a library of motifs. Rooman et al. (1990a) discuss different pruning techniques, and point out the dependence of any technique on the pragmatics of the motif library. In a first experiment, a fixed number (four) of general motifs is sought after, for each fragment length from four through seven. These motifs are correlated with four standard secondary structures. The four discovered classes are used in the sequence–structure analysis (Rooman et al., 1990b) discussed in the previous section. In a second experiment, Rooman et al. (1990a) cluster heptamers and only retain motifs which contain, on average, 50 members. Some of these motifs are subjected to a structure–sequence analysis (see next section).

Unger et al. (1989) first perform an optimal alignment of two fragments in 3D before their comparison, using a best-molecular-fit routine (e.g., Kabsch, 1976). Virtual backbone, described by Cα positions of residues, are clustered using a k-nearest neighbor algorithm. Whereas Rooman et al. (1990a) use the DRMS metric, based on intra–fragment distances, Unger et al. first optimally align the structures and use the "aligned" root-mean-square deviation (ARMS) between the two aligned point sets[2] (Cohen & Sternberg, 1980) as a measure of similarity. Various criteria are used to evaluate the proposed clustering, including the production of a reasonable number of discovered motifs, and a reasonable goodness of fit to the training fragments.

Nussinov and Wolfson (1991) use large–scale structural comparisons to discover protein structure motifs. Unlike the approaches surveyed above, a protein is not divided into contiguous training fragments; rather, two proteins are compared in their entirety to identify atoms that are superposable in 3D. A set of such points forms a structural motif. A technique from computer vision known as *geometric hashing* is used to compute the geometric alignment parameters which produce the most superposable atoms. Two interesting features of this method are that motifs need not refer to contiguous sequential regions, and that they can contain arbitrarily many components.

Table 1 summarizes several structure motif discovery results. The first column indicates the method for computing similarity between two fragments; whether they are first

Table 1. Structure motif discovery results (—: not reported, †: per protein pair).

Paper	similarity	fragment length	# fragments	# motifs
Zhang et al., 1993	invariant	7	6,422+6,242	6
Prestrelski et al., 1992	invariant	8	2,378	113
Hunter & States, 1991	standardized	5/6	9,556/9,491	27/—
Rooman et al., 1990a	invariant	4–7	13,000	4
Rooman et al., 1990a	invariant	7	13,000	> 10
Onizuka et al., 1993	invariant	5–129	—	16
Nussinov & Wolfson, 1991	alignment	full protein	2	1†
Unger et al., 1989	alignment	6	13,000	103
Matsuo & Kanehisa, 1993	alignment	7	15,320	37

standardized, aligned during comparison, or whether features invariant under Euclidean transformations are used. The second column indicates the training fragment size; there is some consensus on a length from six to eight. The third column indicates the number of training fragments used. Finally, the last column indicates the number of retained motifs. It is apparent that there is little agreement on the optimal motif library size. As we have seen, however, different research has different pragmatic goals for the discovered motif libraries.

2.3. Structure–sequence motifs

Structure–sequence motifs assign both sequence and 3D coordinate information to residues. This motif type is different from the sequence–structure motif in that the motif itself must have an explicit 3D structure. The sequence–structure motifs described in the previous section do not fall into the category of structure–sequence motifs because the 3D structure is only implicit in the association of a residue with a structure identifier.

Structure–sequence motifs are currently receiving a great deal of attention in the molecular biology literature. The "inverse" structure prediction problem, where a sequence is predicted for a given structure, relies on a compact library of structure–sequence motifs. The sequence portion of these motifs can be a conserved or property sequence, and may be probabilistic in that propensities for different amino acids at each sequence position are indicated. A motif, which may be noncontiguous, is "threaded" by the amino acid sequence of unknown structure, with an evaluation function judging the goodness of fit of the input sequence to the motif sequence for a particular threading (Jones et al., 1992; Ponder & Richards, 1987). One attraction of this approach is that the number of possible protein structure classes may not be very large — on the order of one thousand (Chothia, 1992) — and that an exhaustive search over a compact library of structure–sequence motifs may be feasible.

Many researchers concerned with the discovery of structure motifs (see previous section) have attempted to generalize their techniques to produce structure–sequence motifs. This involves an analysis of the sequences within a discovered structure class. Unger et al. (1989) tabulate statistics on the frequency of each amino acid types at every position

in a structure motif, producing a consensus sequence motif. Preliminary results indicate that the local 3D structure of a fragment can sometimes be predicted by assignment of the fragment to a motif based on these frequency tables. Zhang et al. (1993) also tabulate the relative frequency of each of the amino acids for the central residue in their six discovered structure motifs. Finally, Rooman et al. (1990b) obtain a property sequence motif for discovered heptamer structures. These sequences, however, are only weakly specific and are not likely to be predictive of the associated structure.

Each of these three simple extensions to structure motif discovery are restrictive in that a *single* property or conserved sequence must be associated with a structure; it is not possible to represent a structure which is associated with a disjunction of sequences. This is a problem because the consensus sequence of a discovered structure motif will tend towards the flat distribution. The discovery method presented in Section 3 of this paper removes this restriction. Also, most structure–sequence work reflects a pragmatic bias to the prediction of structure from known sequence. As we have seen, an equally useful task is the prediction of sequence from hypothetical or known structure.

2.4. *Discussion*

Table 2 classifies the protein motif discovery work discussed above according to three dimensions. Column 2 indicates the motif type. Columns 3 and 4 indicate the semantic theory which dictates the meaning of a motif, and hence the motif-fragment relationship. In a *model-theoretic* semantics, the extension of a motif is a set — the set of all fragments with the same properties and relationships as the motif. In a *probabilistic* framework, a motif also denotes a set, but here the elements of the set have a probability of occurrence. A *similarity* semantics can be given in two ways, one where the motif is assigned a distance threshold δ (as in the work of Rooman et al., 1990a), another where there is no bound and the motif denotes something similar to a "fuzzy" set. Researchers in protein motif discovery are often not clear about which semantics is intended. Table 3 summarizes the three main semantic theories of protein motifs. For example (row 1), if a motif X has a similarity (δ) semantics, then a fragment Y must meet the condition $d(X, Y) \leq \delta$, where δ is a researcher-specified constant.

Table 2 shows that, in the work surveyed, protein motifs have not been given a common semantics. Structure motifs have usually been represented using prototypes that have a similarity semantics, whereas sequence motifs have usually been represented by logical definitions with a model-theoretic semantics. Structure–sequence motifs often inherit a confused dual semantics.

Most protein motif discovery work uses a restricted semantics for structure motifs, in that a motif can only subsume a fragment of the same length. This leaves little flexibility for reasoning about relative generality of motifs, an important capability (Rooman & Wodak, 1991). Unger et al. (1989) note that the ARMS (and DRMS) measures of structural similarity can give counterintuitive results, not necessarily reflecting topological or qualitative similarities. The following section proposes a representation that has a model-theoretic semantics for both structure and sequence motifs and addresses many of the issues raised in this survey.

Table 2. A comparison of different protein motif representations (—: not applicable to this type of motif).

Paper	Motif type	Semantic Theory	
		Sequence	Structure
Taylor, 1986	sequence	model	—
Smith & Smith, 1990	sequence	similarity	—
Rooman & Wodak, 1988	seq-struct	model	—
Rooman & Wodak, 1991	seq-struct	model	—
Rooman et al., 1990b	seq-struct	model	—
Hunter & States, 1991	struct	—	probability
Rooman et al., 1990a	struct	—	similarity
Nussinov & Wolfson, 1991	struct	—	similarity
Prestrelski et al., 1992	struct	—	similarity
Onizuka et al., 1993	struct	—	similarity
Matsuo & Kanehisa, 1993	struct	—	similarity
Unger et al., 1989	struct-seq	probability	similarity
Blundell et al., 1987	struct-seq	model	similarity
Sali & Blundell, 1990	struct-seq	similarity	similarity
Zhang et al., 1993	struct-seq	probability	similarity
Conklin et al., 1993	struct-seq	model	model

Table 3. The relationship between a motif X and a fragment Y in various semantic theories.

Semantic theory	X	# values for truth	notation
similarity (δ)	prototype	2	$d(X,Y) \leq \delta$
similarity	prototype	—	$d(X,Y)$
probabilistic	probability distribution	many	$p(Y\|X)$
model	logical sentence	2	$\models Y \Rightarrow X$

3. Discovery in a Spatial Description Logic

The first part of this paper presented the important application of protein motif discovery, and surveyed a number of existing approaches. Many issues were raised and discussed, including the semantics given to motifs, the motif evaluation mechanism employed, and the pragmatics of discovered motifs. This part of the paper will present a new approach to protein motif representation and discovery which is based on knowledge representation ideas from description logics and machine learning principles of structured concept formation.

Concept formation systems construct a hierarchical organization of intensional concept definitions (Gennari et al., 1989), and should therefore be based on a theory of concept generality and subsumption. Description logics (Nebel, 1990) are a restricted first-order formalism with specialized inference rules for detecting extensional subsumption. Description logics provide an elegant underlying formalism for concept formation work, useful for both background knowledge and discovered concepts. Section 3.1 briefly reviews general description logics, and presents $\mathcal{SDL}$, a *spatial* description logic specifically tailored to reasoning about structured concepts.

A *structured object* is composed of *parts* along with defined *relations* among these parts. A structured object may be *composite*, recursively comprising other structured objects as parts, or *atomic*.[3] The *level* of a structured object is defined inductively as follows. The level of an atom is 0. The level of a composite object is one greater than the maximum of each part level. Although $\mathcal{SDL}$ is capable of describing multilevel objects, only level-0 and level-1 objects will be encountered in this paper. Supervised concept learning with structured objects falls into the general area of *relational learning*, which is being explored currently by inductive logic programming researchers (Muggleton, 1992). The incremental conceptual clustering of structured objects is addressed by the field of *structured concept formation* — see Thompson and Langley (1991) and Conklin (1995) for reviews of the field.

Section 3.2 presents a structured concept formation system which discovers $\mathcal{SDL}$ concepts. The structural similarity of two fragments is measured according to the number of parts participating in a relation-preserving correspondence between them. A match between two compellingly similar fragments is used to build a generalized motif, which is then classified into an evolving concept taxonomy. The structured concept formation procedure was applied to a large database of fragments drawn from different structural classes of proteins. Section 4 presents these results.

Concept discovery is a highly underconstrained task, and it is essential that some form of selective acquisition and retention (Markovitch & Scott, 1993) of discovered concepts be performed. The concept formation system described in Section 3.2 strives to create a taxonomy which is a fast information retrieval system for fragments. In Section 4, structure–sequence motifs are evaluated according to their predictive power, that is, the ability of the sequence of a motif to predict its associated structure. Examples of high-quality, predictive discovered motifs are presented in Section 4.

3.1. A Spatial Description Logic

Description logics are a frame–based representation scheme which make a clear division between concepts (called the *terminology*) and instances of those concepts (described using *assertions*). A terminology is created using the `defconcept` and `defprimconcept` statements which associate concept names with concept terms. The concept term `any` is predefined. Concept names may not be defined more than once in a terminology. Concept terms are constructed from other concept terms using concept constructors such as conjunction, negation and disjunction.

The central reasoning method in description logics is reasoning about *subsumption* of concepts. In any consistent model of a terminology, each concept defines an *extension*: the set of objects in a domain of interpretation that are instances of the concept. One concept C *extensionally subsumes* another concept D in a terminology $\mathcal{T}$, denoted $\mathcal{T} \models C \succeq D$ or more simply $C \succeq_{\mathcal{T}} D$, if its extension is a superset of the other's in all possible models of $\mathcal{T}$. Two concepts C and D are *extensionally equivalent*, denoted $C \equiv_{\mathcal{T}} D$, if they co-subsume each other. The subsumption relation induces a *concept taxonomy*. This is a lattice denoting the partial order of subsumption between concept names. For example, the relation `amino-acid` $\succeq_{\mathcal{T}}$ `Glycine` is depicted by the concept

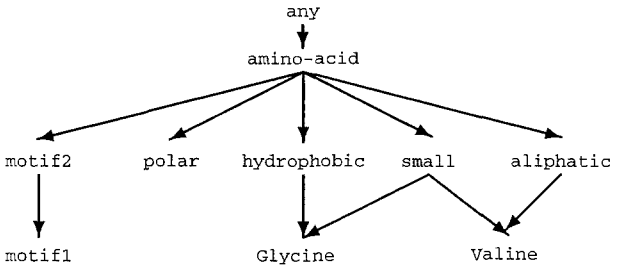


Figure 3. A concept taxonomy depicting a portion of Taylor's (1986) domain theory.

taxonomy Figure 3. This small taxonomy illustrates a portion of Taylor's (1986) domain theory of amino acids; Taylor's complete classification can be represented in description logic terms. The concept `Glycine`, for example, is defined by the statement

```
defprimconcept Glycine (small and hydrophobic);
```

which states that all glycines are necessarily small and hydrophobic. The taxonomy of Figure 3 also displays a subsumption relationship between two protein motifs `motif2` and `motif1`, as discussed later in this section.

All description logics possess inference rules for detecting subsumption between two concept terms; these procedures compute the validity of the sequent $\mathcal{T} \vdash C \succeq D$ (read "from background knowledge $\mathcal{T}$, it can be inferred that C subsumes D"). One of the main facilities of a description logic is a *classifier*, which places a new concept in its "correct" location in the taxonomy, just below all most specific subsumers (MSS), and just above all most general subsumees (MGS) (Woods, 1991). As a new concept progresses down the taxonomy during the classification process, concepts which are increasingly specific and similar to the new concept are encountered.

Description logics have a method for expressing relationships between objects; these so–called *roles* are restricted to binary relations. In order to facilitate reasoning about structured objects, a *spatial* description logic called $\mathcal{SDL}$ has been crafted. The main addition made by $\mathcal{SDL}$ to standard description logic principles is the *image term*. Image terms are used to represent structured concepts, and are formed by associating a *symbolic image* with a set of *relation* identifiers.

A symbolic image is described by a spatial data structure comprising a set of concept terms with their coordinates in multidimensional space. Informally, a symbolic image is a set of *components*, or term/coordinate pairs. Formally, the abstract data type *Image* is given by the following signature:

$$\begin{aligned} Component &= ConceptTerm \times Coordinate \\ \texttt{empty} &: \emptyset \rightarrow Image \\ \texttt{put} &: Image \times Component \rightarrow Image \\ \texttt{delete} &: Image \times Component \rightarrow Image \end{aligned}$$

Many other operations on images can be defined using these constructors: for example, a `replace` operation which changes a component, a `move` operation which transfers a single part to a new location, an `overlay` operation which "merges" two images, or a `extract` operation which retains only those components of an image lying in a given region of space. The *canonical form* of a symbolic image is given by a sequence of `put` operations. This canonical form will be abbreviated below simply as a set of components.

An *image term* is formed by associating a symbolic image with a set of relation identifiers that are preserved by the image. These relations are analogous to description logic roles, except that they are computed by functions which directly manipulate the symbolic image data structure. Functions that operate on symbolic images, computing n-ary relations, take n-tuples of components as arguments. The $\mathcal{SDL}$ representation language provides a *diagrammatic* rather than a *sentential* representation (Larkin & Simon, 1987) of structured concepts. Diagrammatic representations have a number of nice properties for structured concept representation, including simple and elegant rules for structural subsumption and generalization (Table 4), compactness of description, and an exact structural and geometrical correspondence with objects from the domain. The latter property will be important in Section 4 where one of the motif evaluation criteria is the average "visual" similarity of instances.

The semantics of $\mathcal{SDL}$ is straightforward, deriving from standard description logic semantics, except that image terms have a unique interpretation. The extension of an image term is the set of all things with the mentioned parts in the mentioned relationships (an illustration of image term semantics will be given below). The induced extensional definition of image term subsumption has a *structural* counterpart. One image term $[I, R]$ structurally subsumes another $[J, R]$ if and only if there exists a relational monomorphism (Haralick & Shapiro, 1993) between their canonical forms that also preserves subsumption. Thus $[I, R] \succeq_{\mathcal{T}} [J, R]$ if and only if there exists an injective function f from the components of I to the components of J such that $\text{first}(a) \succeq_{\mathcal{T}} \text{first}(f(a))$ for all components a of I, and for all tuples of components related (not related) in I, f maps them to components in J which are also related (not related). Table 4 lists four additional structural subsumption rules for $\mathcal{SDL}$.

3.1.1. *Protein motif representation in $\mathcal{SDL}$*

The $\mathcal{SDL}$ representation language has been used successfully in other domains of chemistry, for example, to represent hexopyranose sugar configurations (Conklin et al., 1992). An earlier paper (Conklin et al., 1993) showed how $\mathcal{SDL}$ could be used to represent all four protein motif types. The discussion here will be mainly concerned with structure–sequence motifs.

Sequence motifs can easily be represented in $\mathcal{SDL}$ using a 1D coordinate space[4]:

```
type coordinate = (w : integer);
```

Various semantics have been assigned to sequence motifs, particularly in cases where insertion and deletion of residues are allowed. For now we ignore these cases and

Table 4. Four structural subsumption rules for $\mathcal{SDL}$. $\mathcal{T}$ is a terminology, C and D are concept terms, A is a symbolic image, P is an image component, c is a coordinate, and R, R' are relation identifier sets.

1. $\dfrac{}{\vdash C \succeq (C \text{ and } D)}$	The first deduction rule states that any component of a conjunction subsumes that conjunction. The next three rules apply solely to image terms, and follow directly from the definition of image term subsumption given in Section 3.1. The second rule states that deleting one or more components from an image produces a more general image term. The third rule states that a more general image term can be produced by removing relation identifiers from a relation set. Rule 4 states that replacing a part of a component by a subsuming concept term produces a more general image term.
2. $\dfrac{}{\vdash [\texttt{delete}(A, P), R] \succeq [A, R]}$	
3. $\dfrac{R' \subseteq R}{\vdash [A, R'] \succeq [A, R]}$	
4. $\dfrac{\mathcal{T} \vdash C \succeq D}{\mathcal{T} \vdash [\texttt{replace}(A, (D\ c), (C\ c)), R] \succeq [A, R]}$	

assume that sequence motifs preserve the topological or graph–theoretic distance between residues:

```
distance(p,q) = p.w - q.w;
```

The expression `p.w` refers to the solitary `w` dimension of the component `p`. Sequence motifs are constructed by associating this `distance` relation with a sequence.

To represent protein structure motifs it is necessary to use relations that are invariant under Euclidean transformations. Examples include hydrogen bonding, distance ranges, angle ranges, and ternary spatial relationships. To represent structure and structure–sequence motifs we simply place the parts in a 4D space, using the `w` dimension to represent topological order, and the `x`, `y` and `z` dimensions to represent the Cartesian coordinates:

```
type coordinate = (w :integer, x : real, y : real, z : real);
```

As an example of a relation defined over this space, consider the 4–ary Δ (`delta`) relation, which is defined in terms of the VBDA (`vbda`) between a chain of contiguous (`contig`) residues (for brevity, contiguity conditions are omitted for the L, Z, and J relations):

```
contig(p,q) = true if distance(p,q) == 1;
delta(p,q,r,s) = U if
          contig(p,q) and contig(q,r) and contig(r,s) and
          -75 <= vbda(p,q,r,s) <= 15;
delta(p,q,r,s) = L if
           15 <= vbda(p,q,r,s) <= 105;
delta(p,q,r,s) = Z if
           105 <= vbda(p,q,r,s) <= 195;
delta(p,q,r,s) = J if
           195 <= vbda(p,q,r,s) <= 285;
```

This qualitative relation takes on four discrete values — U, L, Z, J — and tracks how a fragment winds through 3D space. It is called Δ because it looks at the change in direction of a fourth part relative to three other parts. The idea of partitioning the space of VBDAs to create structure descriptors was proposed independently by Conklin et al. (1993) and Ring et al. (1992). The particular partitioning above is due to Ring et al., and is based on a statistical analysis of a large number of protein tetramers. It is attractive for protein motifs since repetitions (called *structural sequences*) of L and J identifiers correspond closely to the standard α and β secondary structures, respectively (Levitt & Greer, 1977).

To define a structure–sequence motif, we first create a symbolic image for a fragment from the Protein Data Bank. For example:

```
defimage 5ADH-heptamer-203 (
      (Valine       (1  11.6  14.8  28.1))
      (Glycine      (2  13.9  14.2  30.9))
      (Leucine      (3  16.4  16.7  29.5))
      (Serine       (4  13.6  19.1  29.1))
      (Valine       (5  12.8  19.0  32.8))
      (Isoleucine   (6  16.4  19.8  33.7))
      (Methionine   (7  16.0  23.0  31.7)));
```

The defined identifier `5ADH-heptamer-203` can now be used in constructing a structure-sequence motif:

```
defconcept motif1 (image 5ADH-heptamer-203 (distance delta));
```

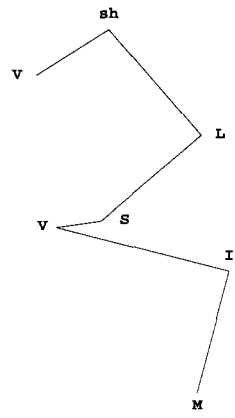
asserting that this symbolic image preserves the topological distance and Δ relations. The `defconcept` construct introduces necessary and sufficient conditions for concept membership. Due to the `defimage` naming facility of $\mathcal{SDL}$, an individual database fragment need only be defined once, and motifs are constructed by associating relations and applying both primitive image operators and generalization operators to that fragment.

Generalization is based on inverting the last three subsumption rules of Table 4. For example, a motif more general than `motif1` (defined above) can be constructed by a single application of the replacement rule:

```
defconcept motif2 (image
   (replace 5ADH-heptamer-203
       (Glycine  (2  13.9  14.2  30.9))
       ((small and hydrophobic)  (2  13.9  14.2  30.9)))
   (distance delta));
```

This motif is more general than `motif1`, since constraints on one of its parts (the glycine) have been weakened (i.e., `(small and hydrophobic)` $\succeq_{\mathcal{T}}$ `Glycine`). The extension of `motif2` includes not only instances with a glycine (appropriately related to other parts of the motif); any residue that is both small and hydrophobic (e.g., alanine, threonine) can be substituted. Thus `motif2` $\succeq_{\mathcal{T}}$ `motif1`, and this is depicted by the concept taxonomy of Figure 3.

The semantics of image terms is given by equivalence to first-order predicate calculus. Figure 4 illustrates this translation for the image term of `motif2`. Images are decomposed



λ x.$\exists^*$ p1 p2 p3 p4 p5 p6 p7

part(x,p1) $\wedge$ ... $\wedge$ part(x,p7) $\wedge$
Valine(p1) $\wedge$
small(p2) $\wedge$ hydrophobic(p2) $\wedge$
Leucine(p3) $\wedge$ Serine(p4) $\wedge$
Valine(p5) $\wedge$ Isoleucine(p6) $\wedge$
Methionine(p7) $\wedge$
distance(p1,p2) = 1 $\wedge$
distance(p1,p3) = 2 $\wedge$...$\wedge$
delta(p1,p2,p3,p4) = L $\wedge$
delta(p2,p3,p4,p5) = L $\wedge$
delta(p3,p4,p5,p6) = L $\wedge$
delta(p4,p5,p6,p7) = L

Figure 4. The logical equivalence semantics for the image term of `motif2` (see text, Section 3.1.1) which preserves the `distance` and `delta` relations. Left: a depiction of the symbolic image. Amino acid abbreviations are standard: amino acid property abbreviations are given in the Appendix. Right: the translation of the image term into a one-variable lambda predicate. For brevity, many of the topological distance terms are omitted from the translation.

according to a fundamental relation (Thompson & Langley, 1991) called `part`. The structural sequence for this helical motif is `LLLL`. Note the use of Haussler's (1989) $\exists^*$ quantifier, which ensures that each quantified variable refers to a distinct object. The extension of a one-variable lambda predicate, with respect to a domain of interpretation, is all objects in the domain for which the predicate is true. Under this semantics, it can be shown that the last three subsumption rules of Table 4 are sound: the proofs are straightforward and are omitted here.

An $\mathcal{SDL}$ concept taxonomy is a lattice structure with concept names as nodes. These concept names are defined using `defconcept`, and will usually refer to image terms. Very general motifs are placed at high levels of the taxonomy. The actual database fragments will be at the leaves of the taxonomy. The taxonomy structure is incrementally revised by a machine discovery procedure, which is the topic of the next section.

3.2. *Discovery in the Spatial Description Logic*

The IMEM (Image MEMory) system (Conklin & Glasgow, 1992) is a similarity–based structured concept formation system which discovers, revises, maintains and organizes an $\mathcal{SDL}$ knowledge base of images. The original purpose of IMEM was as a generalization-

Table 5. The IMEM algorithm. Top–level call: `incorporate(image)`.

```
incorporate(image)
    let M be the most specific subsumers of image
    for each concept C in M do
        (a) place a subsumption link between C and image
        (b) merge(image, C)
merge(image, C)
    for all immediate subsumees D of C such that S(image, D) ≥ τ do
        (a) form a new concept which is a common subsumer of image and D, and give
            it a unique name U
        (b) classify(U)
classify(N)
    if N is equivalent to an existing concept, create a pointer from N to that
       concept
    else place N in the current concept taxonomy just below all most specific
       subsumers, and just above all most general subsumees
```

based memory (Lebowitz, 1987) for efficient retrieval of images (structured objects) for analogical reasoning. Though IMEM can be seen as a conceptual clustering system — grouping instances into classes of high similarity — it is perhaps better viewed as a system for learning by analogy (Winston, 1980). A high-level description of the algorithm is given in Table 5. The IMEM method differs from concept formation methods such as UNIMEM (Lebowitz, 1987) and COBWEB (Fisher, 1987) with respect to the structured hypothesis space used, the rigorous denotation of links in the concept hierarchy, and in the way similarity is measured. IMEM learns by recalling images similar to an input image, and by performing a generalization and a classification step. The concepts of image term subsumption and classification were outlined in the previous section: this section will focus on the similarity-matching and generalization steps.

IMEM exploits a well-known principle of information retrieval: a hierarchy of clusters can balance both the precision and recall of retrieval and improve retrieval efficiency (Salton & Wong, 1978). As Levinson (1985) has noted, a concept taxonomy can be used for close-match retrieval simply by computing the most specific subsumers (MSS) of an input concept, followed by a close-match computation with each instance of each MSS. This close-match retrieval principle is also exploited in the `merge` step (see Table 5), with minor modifications for efficiency purposes; first, only the *immediate* subsumees of each MSS are scanned for similarity, and secondly, generalization only occurs with concepts more similar than a researcher-specified threshold τ.

There are many techniques for measuring the similarity between two structured objects in machine learning (e.g., Falkenhainer et al., 1989; Bisson, 1992; Winston, 1980). Similarity theory is also an active research area in structural chemistry (Johnson & Maggiora, 1990). The similarity of two images in IMEM is primarily *structural.* IMEM was originally devised as a memory for analogical reasoning, and implements the *structure*

mapping principle of analogy (Falkenhainer et al., 1989). The generalization operation in the `merge` step uses a computed structural mapping, which explicitly applies the part deletion subsumption rule (Table 4), and generalizes the matched components with the replacement rule. The use of the relation subset rule will be encountered in Section 4.

3.3. Image similarity

Two image terms C and D are *structurally equivalent* if they have the same relation sets, the same number of parts, and if these parts are identically related. More precisely, there must exist a relational isomorphism which maps the parts of one image to the parts of the other. The *similarity* of two images having the same relation sets can be measured in terms of the number of part deletions needed to bring them into structural equivalence, that is, according to the size of a *partial* relational isomorphism between them. Similarity also recursively takes into account the similarity of corresponding image parts.

My Ph.D. dissertation presents an axiomatic theory of similarity for $\mathcal{SDL}$. A similarity coefficient $S_{\mathcal{T}}$ must obey several properties, including symmetry and monotonicity. It can be shown that the Dice and Tanimoto similarity coefficients, popular in chemical information retrieval systems (Willett, 1990), have these properties. The following proposition, proved in my dissertation, establishes the close relationship between similarity and subsumption in $\mathcal{SDL}$.

PROPOSITION 1 (INDEXING) *Let $\mathcal{T}$ be a terminology, C and D be any two image terms such that $C \succeq_{\mathcal{T}} D$. For any image term I, if $D \succeq_{\mathcal{T}} I$, then $S_{\mathcal{T}}(D, I) \geq S_{\mathcal{T}}(C, I)$. Furthermore, the inequality is strict whenever the relation $C \succeq_{\mathcal{T}} D$ is proper (i.e., $C \not\equiv_{\mathcal{T}} D$).*

The impact of the indexing proposition is that the concept taxonomy can be used to index images for close-match retrieval. The proposition guarantees that as a target image descends a concept taxonomy, fewer and more closely matching source images will be retrieved. Thus retrieval is directed, and need not involve enumeration of images.

3.3.1. Least common subsumers

The computation of a structural similarity valuation between two images C and D necessarily produces a set of partial relational monomorphisms between them. Each one of these functions induces a common subsumer L which, by definition, is structurally equivalent to a subimage of C and a subimage of D. The parts of L are produced by computing a common subsumer of corresponding parts in C and D. If the relational monomorphism is maximal (i.e., not a sub-function of any other) and a least common subsumer is computed for corresponding parts, the image term L will be a least common subsumer of C and D.

To illustrate this, refer to Figure 5, which shows two images C and D (top left and right) and below them a least common subsumer. Although not apparent from the diagrams, the structural sequence of C is `LLJU`, and the structural sequence of D is `LLJJ`. A maximal

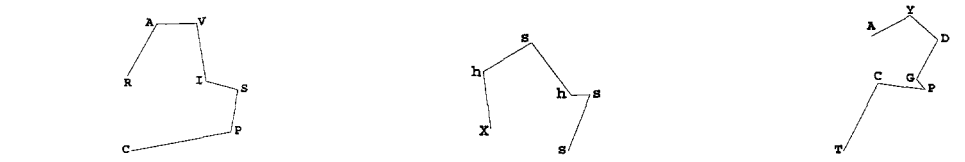


Figure 5. A least common subsumer (middle) of two images (left, right). The two fragments are not structurally equivalent under the `delta` relation; one part must be dropped from each to make them so. See Section 3.3.1 for further details.

relational monomorphism between C and D has six elements. Hence one part must be deleted from both C and D (the cysteine (C) and the threonine (T), respectively) to bring them into structural equivalence. The least common subsumer (bottom) therefore has six parts, each of which are least common subsumers of corresponding parts of C and D. The background knowledge $\mathcal{T}$ used during the least common subsumer computation contains Taylor's domain theory of amino acids.

4. Experimental results

4.1. Heptamer motifs

To evaluate the performance of IMEM on structure–sequence motif discovery, a large database of protein fragments was constructed from 99 proteins. These proteins are a union of those used by Hunter and States (1991) and Rooman and Wodak (1991), and each has a resolution better than 2.5 Å. They also span a variety of distinct protein structural families. A database of 17138 virtual backbone heptamer fragments was created by sliding a window of length 7 over each protein virtual backbone sequence. Taylor's (1986) domain theory of amino acid physicochemical properties was coded as background knowledge in $\mathcal{SDL}$. The additional amino acid classes of hydrogen-bond acceptor and donor (used by King and Sternberg, 1990) were coded. Fragments preserved the Δ and the topological distance relations (see Section 3.1.1).

The IMEM method (Table 5) was applied to the fragment database (its order was randomized), with one small modification: when a new structure–sequence motif was created, a "pure" sequence motif (reversing the third subsumption rule of Table 4) and a "pure" structure motif were generated and classified. The generation of corresponding sequence and structure motifs for every structure–sequence motif produces an efficient indexing structure for fragment classification, and provides an elegant and efficient way to evaluate motif sequence–structure predictivity. For the heptamer experiments, fragments are used to create a motif only if they are structurally equivalent.

The efficiency of classification in $\mathcal{SDL}$ degrades (at worst) linearly with the number of concepts in the taxonomy. Nevertheless, the fragment database is very large, and to reduce overall discovery time concept formation was disabled after 2000 fragments were

Table 6. A sequence–structure contingency table used for the chi-square assessment of a discovered motif (motif 5314 of Table 7).

		Structure	
Sequence	yes	no	Totals
yes	11	2	13
no	62	17063	17125
	73	17065	17138

Table 7. Examples of discovered predictive protein structure–sequence heptamer motifs. Id: a unique identifier for the motif. Amino acid property abbreviations are given in the Appendix. Structural identifiers are defined in Section 3.1.1. M_+: the number of fragments subsumed by the motif; M_-: the number of fragments subsumed by the sequence, but not by the associated structure; $\overline{\text{DRMS}}$: mean DRMS (Section 2.2); σ: standard deviation; χ^2: chi-square value computed from the sequence–structure contingency table for the motif. Table 6 displays the contingency table for motif 5314.

Id	Sequence	Structure	M_+	M_-	$\overline{\text{DRMS}}$	σ	χ^2
5314	(s)(ayn)(adp)(s)(h)(P)(h)	JZLZ	11	2	0.49	0.43	2174.10
4791	(adpt)(adp)(X)(ap)(G)(dp)(adpt)	ZJLJ	5	1	0.82	0.68	681.02
6085	(G)(l)(h)(S)(h)(G)(dp)	ZUZJ	9	0	0.48	0.41	2698.42
3450	(V)(l)(adps)(A)(A)(H)(C)	LJUZ	7	0	0.23	0.07	2136.12
6012	(h)(l)(T)(A)(hs)(H)(C)	LJUZ	9	0	0.22	0.07	2746.76
3105	(X)(adpt)(Q)(hs)(s)(S)(hs)	ZLLU	6	0	0.66	0.48	3019.41
6479	(dp)(X)(h)(X)(dyp)(T)(L)	LLLL	5	0	0.24	0.07	19.53
6711	(t)(s)(h)(adp)(adyps)(A)(hs)	LLLL	5	0	0.14	0.03	19.53
5091	(h)(h)(ayp)(r)(l)(X)(N)	LLLL	8	0	0.29	0.12	31.25
5270	(h)(ap)(A)(l)(X)(c)(h)	LLLL	12	1	0.34	0.23	41.46
5382	(X)(A)(X)(l)(dhp)(s)(l)	LLLL	14	3	0.40	0.18	40.26
3993	(s)(s)(H)(s)(A)(G)(hs)	LLLL	5	0	0.38	0.14	19.53

processed. The remaining fragments were then classified: in this manner accurate statistics on the previously discovered motifs are obtained. About 2500 sequence–structure motifs were discovered by IMEM. Each represents a recurrent association between sequence and structure. This experiment explored whether some of the motifs had a sequence which was predictive of local tertiary structure. To this end, those motifs for which the sequence is probably not predictive were filtered out from the original set.

Each discovered motif was passed through three successive filters. First, the ratio M_+/N — where N is the number of fragments subsumed by the sequence portion of the motif, and M_+ is the number of fragments subsumed by the structure–sequence motif — must be greater than 0.8. This ensures that more than 80% of the instances of the sequence motif have the same structure, and therefore that the sequence may be predictive of the corresponding structure. Second, the mean DRMS of all instances of the motif must be less than 0.95 Å[5]. This is done because it does not necessarily follow that motifs, qualitatively similar according to the Δ relation, are also quantitatively or "visually" similar. This filter ensures that the structure portion of the motif is a good building block. Finally, the third filter applies a chi-square test of statistical independence to the discovery. This is a standard technique for evaluating the significance of a suspected

Table 8. Examples of variable-length discovered motifs.

Id	Sequence	Structure	M_+	M_-
7205	(C)(X)(G)(D)(S)(G)(G)(s)(h)	LLZLZJ	9	1
6791	(N)(S)(l)(s)(X)(X)(dp)(D)(I)(h)(L)(L)(K)(L)(ayp)(dhp)	LLJZLZZZJJJJJ	3	0
5289	(X)(dhp)(dp)(h)(N)(h)(aps)(r)(ayps)	JZJZJJ	4	0
3816	(s)(dp)(V)(A)(X)(h)(h)(hs)	LLLLL	7	0
6664	(ap)(X)(h)(X)(dp)(L)(X)(dyp)(X)(G)(h)(X)	LLLLLLLJZ	5	0

association (Zembowicz & Zytkow, 1992). A 2×2 contingency table for the motif is constructed (see Table 6), and a Bonferroni-adjusted chi-square test at a significance level of 0.01 is applied[6]. This test gives support to the conclusion that the sequence and structure of the motif are significantly correlated and not merely random associations.

If a motif C passes through all three filters, then all motifs D such that $C \succeq_T D$ are not considered. In this manner, the *most general* motifs satisfying all three filters are retained. After application of the filters to the 2500 discovered motifs, 144 motifs were retained. A few of these are presented in Table 7. The mean DRMS for many motifs (e.g., 6479, 6711) is very low. Some motifs (e.g., 3450, 3993) have several specific conserved residues indicated. Some motifs (e.g., 5382) are quite general: this indicates their wider applicability for tertiary-structure prediction. Many of the structural sequences are LLLL — referring to the common helical structure — but several other structural sequences are also exhibited.

4.2. *Variable-length motifs*

Given the apparent success of the heptamer analysis, it is reasonable to ask: do there exist longer predictive motifs, and can these be automatically discovered? To answer these questions, a training set was created by sliding a window of length 20 over all 99 proteins, creating 15747 20-mers. In contrast to the heptamer analysis, here IMEM was permitted to delete parts from fragments when creating least common subsumers. Concept formation was disabled after 1000 fragments were processed. About 300 motifs were discovered, and about half of these passed through the three filters. Several motifs were very long — nearing length 20 — and many shorter motifs were also discovered. Table 8 lists just a few of the discovered variable length motifs. Motif 6791 has 16 parts, with several specific residues indicated. Motif 7205 has a very specific sequence, yet covers nine fragments. It is also quite interesting in other respects. It represents an abstraction of residues 191 through 199 in the family of *serine protease* proteins. This region represents the most highly conserved region in serine proteases, and contains the active-site serine residue 195 (Smith & Smith, 1990).

4.3. Discussion and future work

These results address a key question posed by Unger et al. (1989), showing that it is possible to automatically discover protein structural building blocks that carry sequence specificity. These structural building blocks are encapsulated in structure–sequence motifs and expressed as $\mathcal{SDL}$ concepts. Filters were used to prune the set of discovered motifs. Though the filters were tailored for sequence–structure predictivity, there is no reason why structure–sequence predictivity could not also be explored in this discovery framework. For such an experiment, it will probably be necessary to use long training fragments to allow for a wider distribution of structural sequences.

Discovered motifs could be used for protein tertiary-structure prediction, and increasing the length of discovered motifs may allow us to consider tertiary interactions which are nonlocal in the protein sequence. Interactions between distant residues, and the failing of structure prediction methods to take them into account, is one of the hypothesized reasons for the limited prediction success that has been achieved. In general, with longer motifs, more contiguous residues can be predicted, and less tertiary alignment of predicted portions needs to be performed. It is, however, too restrictive to discover motifs of only one size. An advantage of IMEM is that it can discover variable–length motifs.

The efficiency of the IMEM discovery algorithm depends on its incrementality combined with the use of an evolving concept taxonomy for fast close-match retrieval of concepts. Due to incrementality, a different ordering of the training data may produce quite a different concept taxonomy. Once a motif is discovered and classified, it hides its instances from future fragments which it does not subsume (although note that other motifs may subsume some of the instances). For example, in the experiments above, a sequence motif alone will hide some new fragments from subsumed structure–sequence motifs. Although not always desirable, the effect can be exploited by using sequence motifs suspected to be predictive (Rooman et al., 1989) as background knowledge. Driven by empirical data, the discovery system will elaborate these sequence motifs, and also confirm whether they occur in similar structures.

A visual inspection of several discovered motifs has revealed that pairs of motifs — while not subsuming each other — are often quite similar in their sequences. While they both may be predictive, their least common subsumer may not be. This points to the idea of extending $\mathcal{SDL}$ to deal with disjunction and negation to more concisely express the components of protein motifs. Indeed, parts of Taylor's original domain theory of amino acid types is expressed using disjunction and negation. Work on compiling a more extensive knowledge base of amino acids is in progress.

On a related note, all previous structure–sequence discovery work has assumed that the components of motifs — amino acid identifiers — are devoid of manipulable 3D structure. An extension of previous work is to base the discoveries of a system on the internal spatial structure of the amino acids. There are two approaches to such an extension. One is to code, as background knowledge, the definitions from manual amino acid rotamer classifications (Ponder & Richards, 1987). Another approach is to use IMEM to autonomously discover its own rotamer classes. Both approaches require a knowledge representation, such as $\mathcal{SDL}$, capable of describing multilevel structured

objects. Initial results on level-2 protein motif discovery with IMEM is reported by Conklin et al. (1994). Future work could also focus on using discovered heptamer motifs to parse or "tile" a protein. This may provide the basis for the discovery of recurrent level-3 (super-secondary) structural motifs.

5. Conclusions

The field of molecular biology is rich in applications for machine learning. This paper has presented and surveyed the important application domain of protein motif discovery. A novel representation and discovery scheme was outlined, wherein all types of protein motifs have a model-theoretic semantics, and structured concept formation is used to discover recurrent motifs. The organization of a knowledge base by subsumption guides new training fragments to other similar fragments, which are generalized and classified into an evolving concept taxonomy. Some promising results on a large database of protein fragments were presented.

The main emphasis of this paper was on structure–sequence motifs — patterns of 3D structure with attached sequence information. Until now, approaches to structure–sequence motif discovery were limited to the *a posteriori* construction of a conserved or property sequence motif for instances of a discovered structure motif. These approaches do not represent structure motifs associated with a disjunction of sequences, and will produce overly-general and weakly predictive structure prediction rules. The system presented in this paper is the first to fully integrate the representation and discovery of protein structure and sequence motifs. The discovery of protein motifs is part of a larger research program which is concerned with the automated discovery of associations between molecular topological structure and 3D conformation in large molecular structure databases (Conklin, 1995). Description logics and the machine learning paradigms of structured concept formation and learning by analogy have been influential in this research.

Acknowledgments

I would like to thank Jude Shavlik and the anonymous referees for their valuable comments on this paper. I would also like to thank my Ph.D. advisors Janice Glasgow and Suzanne Fortier for their support of this research. The protein motif display software was developed by Chris Walmsley. Financial support for this research was provided by the Natural Science and Engineering Research Council of Canada.

Appendix

Amino acid property abbreviations

abbreviation	description	amino acids with property
X	any amino acid	all
t	tiny	ACS
h	hydrophobic	AFGHIKLMTVWY
y	hydrophylic	SNDEQR
s	small	ACDGNPSTV
c	charged	DEHKR
v	positive	HKR
n	negative	DE
r	aromatic	FHWY
l	aliphatic	ILV
d	hydrogen-bond donor	WYHTKCSNQR
a	hydrogen-bond acceptor	YTCSDENQ
p	polar	CDEHKNQRSTWY
Z	GLX	QE
B	ASX	DN

Notes

1.
$$\mathrm{DRMS}(X,Y) = \sqrt{\frac{1}{n(n-1)/2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\|X_i - X_j\| - \|Y_i - Y_j\|)^2}$$

where X and Y are two protein fragments of the same length n, indexed by i and j, respectively. $\|\cdot\|$ is the vector norm function, so that $\|a - b\|$ is the Euclidean distance between points a and b.

2.
$$\mathrm{ARMS}(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\|X_i - Y_i\|)^2}$$

where X and Y have been aligned to minimize this expression. Sternberg and Islam (1987) relate DRMS and ARMS for protein fragments by the approximation

$$\mathrm{DRMS}(X,Y) \approx 0.75 \times \mathrm{ARMS}(X,Y) + 0.199$$

3. In the context of description logic theory, Nebel (1990) uses the term *atomic concept* instead of *concept name*. In this paper the term *atomic* refers to objects that are not decomposable into substructures. The term *primitive* – as in Lathrop et al. (1987) and Thompson and Langley (1991) — is not used as it has a specific denotation in description logic theory.
4. An abstract notation is used for type declarations and relation definitions, however the syntax for concept declarations and image terms is similar to that used in the implementation of $\mathcal{SDL}$.
5. Unger et al. (1989) consider two hexamer fragments X and Y to be compellingly similar if $\mathrm{ARMS}(X,Y) < 1.0$. Substituting this value into the approximation given in footnote 2 yields 0.95.

6. In a first pass through the motifs, only those which refute the null hypothesis of statistical independence at the 0.01 level of significance ($\chi^2 > 6.64$) are retained. However, the null hypothesis will be falsely rejected — a Type I error — for 1 in 100 tests. To preserve experiment-wise significance levels, it is therefore necessary to modify the chi-square test with a Bonferroni adjustment (Harris, 1985). This divides the significance level by the number n of tests made, and produces a new critical value for χ^2. If n motifs pass the three filters then these motifs are sent through the third filter again, using an adjusted significance level of $0.01/n$.

References

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

Bisson, G. (1992). Learning in FOL with a similarity measure. In *Proc. AAAI-92*. The MIT Press. 82–87.

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

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature* 357:544–545.

Cohen, F. E. & Sternberg, M. J. E. (1980). On the prediction of protein structure: the significance of the root-mean-square deviation. *Journal of Molecular Biology* 138:321–333.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., & Mornon, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Engineering* 6(4):377–382.

Conklin, D., Fortier, S., Glasgow, J., & Allen, F. (1992). Discovery of spatial concepts in crystallographic databases. *Proceedings of the ML92 Workshop on Machine Discovery*, Aberdeen, Scotland. 111–116.

Conklin, D., Fortier, S., & Glasgow, J. (1993). Representation for discovery of protein motifs. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. AAAI Press. 101–108.

Conklin, D., Fortier, S., & Glasgow, J. (1994). Knowledge discovery of multilevel protein motifs. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press. 96–102.

Conklin, D. & Glasgow, J. (1992). Spatial analogy and subsumption. *Machine Learning: Proceedings of the Ninth International Conference (ML92)*. Morgan Kaufmann. 111–116.

Conklin, D. (1995). *Knowledge Discovery in Molecular Structure Databases*. Ph.D. Dissertation, Queen's University, Canada.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure–mapping engine: algorithm and examples. *Artificial Intelligence* 41:1–63.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2:139–172.

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence* 40:11–61.

Haralick, R. M. & Shapiro, L. G. (1993). *Computer and Robot Vision*, volume 2. Addison-Wesley.

Harris, R. J. (1985). *A Primer of Multivariate Statistics*. Academic Press.

Haussler, D. (1989). Learning conjunctive concepts in structural domains. *Machine Learning* 4:7–40.

Hunter, L. & States, D. J. (1991). Bayesian classification of protein structural elements. *Proceedings of the Seventh IEEE Conference on AI Applications: The Biotechnology Computing Minitrack.*

Johnson, M. A. & Maggiora, G. M., editors. (1990). *Concepts and Applications of Molecular Similarity*. John Wiley & Sons.

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

Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO* 5(4):819–822.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure. *Biopolymers* 22:2577–2637.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica* B32:922–923.

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

Larkin, J. & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* 11:65–99.

Lathrop, R. H., Webster, T. A., & Smith, T. F. (1987). ARIADNE: Pattern–directed inference and hierarchical abstraction in protein structure recognition. *Communications of the ACM* 30:909–921.

Lathrop, R. H., Webster, T. A., Smith, R., Winston, P., & Smith, T. (1993). Integrating AI with sequence analysis. In Hunter, L., editor, *Artificial Intelligence and Molecular Biology*. AAAI/MIT Press. chapter 6.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning* 2:103–138.

Levinson, R. A. (1985). *A Self–Organizing Retrieval System for Graphs*. Ph.D. Dissertation, University of Texas at Austin.

Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *Journal of Molecular Biology* 114:181–239.

Markovitch, S. & Scott, P. D. (1993). Information filtering: Selection mechanisms in learning systems. *Machine Learning* 10:113–151.

Matsuo, Y. & Kanehisa, M. (1993). An approach to systematic detection of protein structural motifs. *Computer Applications in the Biosciences* 9:153–159.

Muggleton, S., editor. (1992). *Inductive Logic Programming*. Academic Press.

Nebel, B. (1990). *Reasoning and Revision in Hybrid Representation Systems*. Springer–Verlag.

Nussinov, R. & Wolfson, H. J. (1991). Efficient detection of three–dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings National Academy Science USA* 88:10495–10499.

Onizuka, K., Ishikawa, M., Wong, S. T. C., & Asai, K. (1993). A multi–level description scheme of protein conformation. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. AAAI Press. 301–309.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology* 193:775–791.

Prestrelski, S. J., Williams Jr., A. L., & Liebman, M. N. (1992). Generation of a substructure library for the description and classification of protein secondary structure. i. overview of the methods and results. *PROTEINS: Structure, Function, and Genetics* 14:430–439.

Ring, C. S., Kneller, D. G., Langridge, R., & Cohen, F. E. (1992). Taxonomy and conformational analysis of loops in proteins. *Journal of Molecular Biology* 224:685–699.

Rooman, M. J., Wodak, S. J., & Thornton, J. M. (1989). Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Engineering* 3(1):23–27.

Rooman, M. J., Rodriguez, J., & Wodak, S. J. (1990a). Automatic definition of recurrent local structure motifs in proteins. *Journal of Molecular Biology* 213:327–336.

Rooman, M. J., Rodriguez, J., & Wodak, S. J. (1990b). Relations between protein sequence and structure and their significance. *Journal of Molecular Biology* 213:337–350.

Rooman, M. J. & Wodak, S. J. (1988). Identification of predictive sequence motifs limited by protein database size. *Nature* 335:45–49.

Rooman, M. J. & Wodak, S. J. (1991). Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. *PROTEINS: Structure, Function, and Genetics* 9:69–78.

Sali, A. & Blundell, T. (1990). Definition of general topological equivalence in protein structures. *Journal of Molecular Biology* 212:403–428.

Salton, G. & Wong, A. (1978). Generation and search of clustered files. *ACM Transactions Database Systems* 3(4):321–346.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*. Springer–Verlag.

Smith, R. F. & Smith, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proceedgins National Academy Science* 87:118–122.

Sternberg, M. J. E. & Islam, S. A. (1990). Local protein sequence similarity does not imply a structural relationship. *Protein Engineering* 4:125–131.

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

Thompson, K. & Langley, P. (1991). Concept formation in structured domains. In Fisher, D. H., Pazzani, M., and Langley, P., editors, *Concept Formation: Knowledge and Experience in Unsupervised Learning*. Morgan Kaufmann. 127–161.

Thornton, J. M. & Gardner, S. P. (1989). Protein motifs and data-base searching. *Trends in Biochemical Science* 14:300–304.

Unger, R., Harel, D., Wherland, S., & Sussman, J. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 5:355–373.

Willett, P. (1990). Algorithms for the calculation of similarity in chemical structure databases. In Johnson, M. A. and Maggiora, G. M., editors, *Concepts and applications of molecular similarity*. John Wiley & Sons. chapter 3.

Winston, P. H. (1980). Learning and reasoning by analogy. *Communications of the ACM* 23:689–703.

Woods, W. (1991). Understanding subsumption and taxonomy: A framework for progress. In Sowa, J. F., editor, *Principles of Semantic Networks*. Morgan–Kaufmann. 45–94.

Zembowicz, R. & Zytkow, J. M. (1992). Discovery of regularities in databases. *Proceedings of the ML92 Workshop on Machine Discovery*, Aberdeen, Scotland. 18–27.

Zhang, X., Fetrow, J. S., Rennie, W. A., Waltz, D. L., & Berg, G. (1993). Automated derivation of substructures yields novel structural building blocks in globular proteins. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. Bethesda, MD: AAAI Press. 438–446.

Received October 1, 1993
Accepted October 26, 1994
Final Manuscript November 10, 1994