From: ISMB-93 Proceedings. Copyright © 1993, AAAI (www.aaai.org). All rights reserved.

Representation for Discovery of Protein Motifs

Darrell Conklin[†], Suzanne Fortier[‡][†], Janice Glasgow[†]
Departments of Computing and Information Science[†] and Chemistry[‡]
Queen's University
Kingston, Ontario, Canada K7L 3N6
conklin@qucis.queensu.ca

Abstract

There are several dimensions and levels of complexity in which information on protein motifs may be available. For example, one-dimensional sequence motifs may be associated with secondary structure identifiers. Alternatively, three-dimensional information on polypeptide segments may be used to induce prototypical three-dimensional structure templates. This paper surveys various representations encountered in the protein motif discovery literature. Many of the representations are based on incompatible semantics, making difficult the comparison and combination of previous results. To make better use of machine learning techniques and to provide for an integrated knowledge representation framework, a general representation language — in which all types of motifs can be encoded and given a uniform semantics — is required. In this paper we propose such a model, called a *spatial description logic*, and present a machine learning approach based on the model.

Introduction

A task of growing importance in the management of biochemical data is the ability to perform generalization and abstraction over large sets of related observations. Abstractions offer a form of conceptual aggregation, modelling, and explanation of observations, a method for practical data compression, and also serve to index large databases for efficient information retrieval. There exist recurrent patterns and rules of structural biochemistry hidden in the Protein Data Bank (Bernstein et al., 1977); knowledge discovery techniques can help to uncover these patterns and rules. Generalized patterns can facilitate information retrieval and data incorporation, providing a conceptual framework in which new structural data can be related. They can also be used for prediction and anticipation of molecular conformation, since they often represent common 3D structural features.

A *protein motif* is an abstraction of some observed pattern of amino acid residues. Protein motifs can be roughly classified into four categories. **Sequence** motifs are linear strings of residues with an implicit topological ordering. **Sequence-structure** motifs are sequence motifs with secondary structure identifiers attached to one or more residues in the motif. **Structure** motifs are 3D structural objects, described by positions of residue objects in 3D Euclidian space. Apart from a topological linear ordering on the residues, structure motifs are free of sequence information. Finally, **structure-sequence** motifs are combined 1D-3D structures that associate sequence information with a structure motif. Figure 1 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.

This paper has two objectives. The first is to provide a survey of previous research on protein motif discovery according to the above categorization. The second aim is to present a new approach to protein motif representation and discovery, which is based on knowledge representation ideas of description logics and machine learning principles of structured concept formation. In the proposed representation language, sequence and structure motifs have a uniform model-theoretic semantics. The processes of generalization and structured concept formation are presented. The paper concludes with a discussion of the role of protein motifs in molecular scene analysis.

Discovery of protein motifs

Research in protein motif discovery generally falls into the area of unsupervised machine learning; a machine is given a set of observations (i.e., a set of unclassified protein fragments) and must find a clustering of these observations along with a description of each cluster (i.e., a motif describing the fragments in the cluster). A rigorous mathematical semantics for a motif is necessary in order to determine whether an observation is an instance of a motif. There has been a considerable amount of research on machine discovery of protein

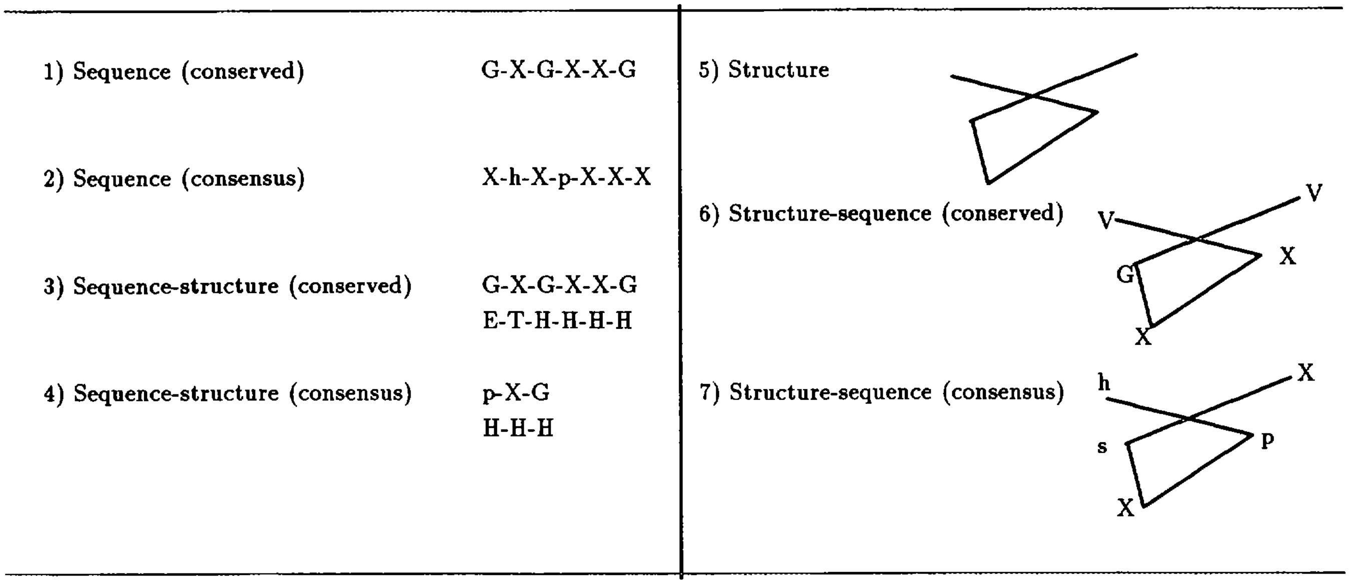
1) Sequence (conserved)
G-X-G-X-X-G
2) Sequence (consensus)
X-h-X-p-X-X-X
3) Sequence-structure (conserved)
G-X-G-X-X-G
E-T-H-H-H-H
4) Sequence-structure (consensus)
p-X-G
H-H-H
5) Structure
6) Structure-sequence (conserved)
V
V
X
G
X
7) Structure-sequence (consensus)
h
X
s
p
X

Figure 1: 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.

motifs. Below we describe and compare a handful of methods in terms of their representation theory, the type of motif under consideration, and the semantics given to motifs.

Structure motifs

Hunter and States (1991) apply Bayesian classification techniques to protein structure motif discovery. This method produces clusterings which are evaluated according to Bayes' formula with a prior distribution favouring 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.

Rooman et al. (1990a) use an agglomerative numerical clustering technique to discover structure motifs, which are represented by a *prototypical* fragment. In contrast to the Hunter and States approach, only the Cα position is used as a descriptor for the position of the amino acid in Euclidian space. Thus amino acids are represented by *point* rather than *line* data. Similarity between fragments is measured by using an RMS metric of the inter-Cα distances. A fragment is an instance of a class by virtue of being within an RMS distance threshold from the prototypical motif of that class.

Sequence and sequence-structure motifs

Protein sequence motifs are the most commonly encountered motif type in the molecular biology literature. There is, for example, an extensive literature on the comparison of sequence motifs: see Lathrop (Lathrop et al., 1993) for a good survey of this work. Sequence motifs are discovered from a maximal alignment of one or more protein sequences, and the abstraction of residues at aligned positions. 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 and Islam, 1990), hence the need for abstraction at alignment positions (e.g., abstracting Ala and Phe to X) to produce a sequence motif. Taylor (1986) uses a more general abstraction scheme; amino acids are classified into nondisjoint groups based on physicochemical properties such as hydrophobicity and polarity which are expected to have an influence on protein folding.

The PIMA method of Smith and Smith (1990) (also see (Lathrop et al., 1993)) discovers consensus sequence motifs which cover an entire set of sequences from functionally related proteins. A physicochemical classification of amino acids somewhat different from Taylor's (1986) is employed. Discovered patterns can contain special "gap" identifiers, which can function as 0 or 1 residues. PIMA is a supervised learning method, since the class of related sequences (e.g., from the protein kinase family) is delimited prior to discovery of a covering sequence motif.

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. Thornton and Gardner (1989) refer to these motifs as "structure-related sequence motifs". For example, Rooman et al. (1989) associate with each amino acid in a motif a standard secondary structure identifier (e.g., Figure 1, motif 3). It is demonstrated that there exist associations for which sequence templates reliably characterize secondary structure. Rooman et al. (1990b) report a study very similar to previous work (Rooman et al., 1989), but instead replace the structure identifiers with identifiers discovered by the previously discussed structure motif analysis (Rooman et al., 1990a).

Harris et al. (1992) use a sequence alignment procedure to discover similarities and clusters of protein sequence regions. Roughly 10, 000 classes of variable-length sequences were uncovered. Their method does not find a comprehensible motif description for discovered classes, therefore it does not address the issue of sequence motif representation.

Two other techniques for protein sequence-structure motif representation should be noted, although they are not concerned explicitly with motif discovery. Cohen et al. (1986) describe the PLANS system which uses a rigorous algebraic notation to describe sequence motifs. The ARIADNE system of Lathrop et al. (1987) represents a protein as a hierarchy of sequence patterns. A pattern can be composite — recursively referring to embedded sequence motifs (e.g., a mononucleotide binding fold pattern) — or primitive (e.g., a residue identifier). The representation described in Section 3 incorporates aspects of both these techniques.

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 are classified by Thornton and Gardner (1989) as "sequence-related structure motifs". They do not fall into our category of structure-sequence motifs because the 3D structure is only implicit in the association of a residue with a structure identifier.

Unger et al. (1989) report an experiment in structure-sequence motif discovery. Hexamers, described by Cα positions of residues, are clustered using a k-nearest neighbor algorithm. As in (Rooman et al., 1990a), similarity between hexamers is measured according to a distance metric. However, similar to Hunter and States' (1991) work, the structures are first aligned using a best molecular fit routine, and absolute coordinates rather than intra-motif distances are compared. Statistics were tabulated on frequency of each amino acid types at every position in a structure motif. Preliminary results indicated 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.

Blundell et al. (1987) present an interesting approach for structure-sequence motif representation and discovery. A common structural core for a set of proteins from a homologous family is first constructed. The sequences corresponding to this core for each training protein are then aligned. Conserved residues are retained and assigned to the common structure motif. A notable difference between this and other structure motif work is that the technique does not require a training set of fragments of fixed size. Blundell's group is concerned with *knowledge based protein modelling*: protein structure prediction, where the main source of information comes from exploration and abstraction of known structures. In this spirit, Sali and Blundell (1990) develop an elaborate scheme for the comparison of protein structures. The results of a comparison form a "generalized protein," which can be used in predicting 3D conformation of the sequence of the unknown. Similar to the work of Lathrop et al. (1987), proteins are described by a hierarchy, with each level being a sequence of typed elements. Elements of fragments are represented by a host of computed *properties*, rather than by a single identifier. Attributes of fragment elements can refer to other elements in the sequence, thus representing binary relationships such as hydrogen bonding between elements. Each of these properties and relationships is weighted, and contributes to an overall weighted distance metric.

Discussion

Table 1 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, a motif denotes 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 2 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 sufficiency condition $d(X, Y) \leq \delta$.

Table 1 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. Unger et al. (1989) summarize some problems with prototypical structure motifs and RMS measures of similarity. The main problem is that two fragments cannot be compared unless they are of the same length. Also, the measure does not necessarily reflect topological and geometric properties of the motifs. The following section

Table 1: A comparison of different protein motif representations.

Author	Motif type	Semantic Theory	
		Sequence	Structure
Hunter and States (1991)	struct	n/a	probability
Rooman et al. (1990a)	struct	n/a	similarity
Taylor (1986)	sequence	model	n/a
Smith and Smith (1990)	sequence	similarity	n/a
Cohen et al. (1986)	sequence	model	n/a
Rooman et al. (1989)	seq-struct	model	n/a
Rooman et al. (1990b)	seq-struct	model	n/a
Lathrop et al. (1987)	seq-struct	similarity	n/a
Unger et al. (1989)	struct-seq	probability	similarity
Blundell et al. (1987)	struct-seq	model	similarity
Sali and Blundell (1990)	struct-seq	similarity	similarity
Conklin et al. (this paper, Section 3)	struct-seq	model	model

Table 2: The relationship between a motif X and a fragment Y in various semantic theories.

Semantic theory	# truth values	X	# values for truth	notation
similarity (δ)	1	prototype	2	$d(X,Y) \leq \delta$
similarity	many	prototype	n/a	$d(X,Y)$
probabilistic	1	probability dist.	many	$p(Y\|X)$
model	1	logical sentence	2	$\models Y \Rightarrow X$

proposes a representation that has a model-theoretic semantics for both structure and sequence motifs.

A Spatial Description Logic

The representation we use for protein motifs is based on terminological or description logics (Nebel, 1990). 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 concept introductions 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. Disjointness restrictions between two concept names state that no object can simultaneously be an instance of both concepts. A small description logic program is given in Figure 2. It defines seven primitive concepts — a statement `defprimconcept x y` asserts "all instances of the concept `x` are necessarily instances of the concept `y`". The program also declares that glycines and valines are disjoint.[1]

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 simply $C \succeq_{\mathcal{T}} D$, if its extension is a superset of the other's in all possible models of $\mathcal{T}$. The subsumption relation induces a *concept taxonomy*. This is a lattice denoting the partial order of subsumption between concept names. For example, `amino-acid` $\succeq_{\mathcal{T}}$ `Glycine` in the terminology of Figure 2, and this is depicted by the associated concept taxonomy. The taxonomy also displays a subsumption relationship between two protein motifs `motif2` and `motif1`, as discussed later in the paper.

Description logics have a technique for expressing relationships between objects; these so–called *roles* are restricted to binary relations. In order to facilitate reasoning about structured objects, we have crafted a *spatial* description logic called $\mathcal{SDL}$. This is a description logic specifically tailored for efficient representation and classification of structured objects. The main addition made by $\mathcal{SDL}$ to standard description logic principles is the *symbolic image*. A symbolic image is described by a spatial data structure comprising a set of concept terms with their coordinates in multidimensional space. These components can be *composite*

[1] This small example illustrates a portion of Taylor's (1986) classification of amino acids; the complete classification can be represented in description logic terms.

```
defprimconcept amino-acid any;
defprimconcept polar amino-acid;
defprimconcept hydrophobic amino-acid;
defprimconcept small amino-acid;
defprimconcept aliphatic amino-acid;
defprimconcept Glycine (small and hydrophobic);
defprimconcept Valine (aliphatic and small);
disjoint       Glycine Valine;
```

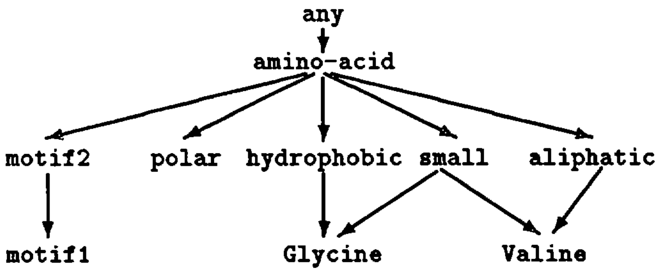

Figure 2: Left : A description logic program; right: its concept taxonomy.

(referencing other subimages), or *atomic.*[2]

Informally, an image is a set of *components*, or term/coordinate pairs. Formally, the abstract data type *Image* of images is given by the following signature:

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

Many other operations on images can be defined using these constructors: for example, a `replace` operation which replaces a component, a `move` operation which transfers a single part to a new location, or an `overlay` operation which "merges" two images:

$$\begin{aligned} \texttt{replace} : Image \times Cmpnt \times Cmpnt \rightarrow Image \\ \texttt{move} : Image \times Cmpnt \times Co\ddot{o}rdinate \rightarrow Image \\ \texttt{overlay} : Image \times Image \rightarrow Image \end{aligned}$$

An *image term* is formed by associating a symbolic image with a set of relations 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 take an image and a tuple of components as arguments; formally,

$$Image \times Component^n$$

is the domain of a function for an n-ary relation r. The symbolic image representation is derived from research on pictorial database systems (Chang and Lee, 1991) and computational imagery (Glasgow and Papadias, 1992).

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. This *extensional* definition of subsumption has a *structural* counterpart. One image term I structurally subsumes another J if and only if there exists a relational monomorphism (Haralick and Shapiro, 1993) between them that also preserves subsumption. Thus $I \succeq_T J$ if and only if there exists a one-one function f from the parts of image I to the parts of image J such that $p \succeq_T f(p)$, and for all parts related (not related) in I, f maps them to parts in J which are also related (not related). The notion of subsumption as a relational monomorphism between images also provides a basis for image structural similarity, which is defined in terms of the number of part deletions needed to bring them into equivalence (Conklin and Glasgow, 1992). A full description of the syntax and semantics of $\mathcal{SDL}$ will appear elsewhere.

This representation logic has been used successfully in other domains of chemistry, for example, to represent hexopyranose sugar configurations (Conklin et al., 1992) and six-member ring conformations. The following subsections will demonstrate how $\mathcal{SDL}$ can represent all of the types of protein motifs discussed in the previous section.

Sequence and sequence-structure motifs

Sequence motifs can easily be represented in $\mathcal{SDL}$ using using a 1D coordinate space[3]:

```
type coordinate = (w : int).
```

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 assume that sequence motifs preserve the graph-theoretic distance between residues:

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

The expression `I.p.w` refers to the solitary `w` dimension of the component `p` of image `I`. Sequence motifs are constructed by associating this `distance` relation with a sequence; for example, the sequence motif "p-X-G"

[2] Nebel (1990) uses the term *atomic concept* instead of *concept name*. We reserve the term *atomic* to apply to objects that are not decomposable into substructures. We do not use the term ***primitive*** (as in (Lathrop et al., 1987)) as it has a specific denotation in description logic theory.

[3] An abstract notation is used for type declarations and relation definitions, however the syntax for concept terms and image terms is similar to that used in the implementation of $\mathcal{SDL}$.

can be represented by the following declaration which associates a name **seqmotif1** with an image term:

```
defconcept seqmotif1 (image
   (  [polar,[1]]
      [amino-acid,[2]]
      [Glycine,[3]])
   [distance]);
```

It is also possible in $\mathcal{SDL}$ to allow for deletions of residues in fragments subsumed by a motif; we simply encode a range constraint into a relation associated with an image. Using this concept, it can be demonstrated that $\mathcal{SDL}$ can represent the types of hierarchical sequence patterns used in ARIADNE (Lathrop et al., 1987).

Sequence-structure motifs can be represented by a a particular secondary structure identifier with a sequence motif (e.g., motif 4 in Figure 1). In the more general case, each residue in a motif can be part of a different secondary structure (e.g., motif 3 in Figure 1). In $\mathcal{SDL}$ it is possible to represent both types of motifs. For example, the first type of sequence-structure motif can be represented by the declarations:

```
defprimconcept Helix any;
defconcept seqmotif2 (seqmotif1 and Helix);
```

The first declaration defines the primitive concept **Helix**, and the second associates this secondary structure identifier with the previously defined sequence motif **seqmotif1**.

Structure and structure-sequence motifs

A symbolic image alone does not have an extensional semantics in $\mathcal{SDL}$; it is only by associating an image with relations, resulting in an image term, that it takes on meaning as a concept. To represent protein structure motifs it is necessary to use relations that are invariant under rotation transformations. Examples include distance ranges (Willett, 1990), angle ranges, and ternary spatial relationships. Relations can be arbitrarily complex, and of any arity.

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 : int, x : real, y : real, z : real).
```

As an example of a relation defined over this space, consider the simple Δ (**delta**) relation, which is defined in terms of the sign (**sgn**) of the torsional angle (**tau**) between a chain of connected (**conn**) residues (defined by Cα positions):

```
conn(I,p,q)  = true if distance(I,p,q) == 1.
delta(I,p,q,r,s) = sgn(tau(p,q,r,s))
       if (conn(I,p,q) and
           conn(I,q,r) and
           conn(I,r,s)).
```

This relation takes on two discrete values, and tracks how a fragment turns — left or right — through 3D space. It is easy to modify Δ to partition the 2π radian space of turns in different ways. For example, a modified Δ relation has been used to define the various conformations of six-member rings (results forthcoming).

To define a structure-sequence motif, we first declare and name a fragment from the Protein Data Bank. For example:

```
defimage 5ADH-hexamer-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]]);
```

The defined identifier can now be used in constructing a structure-sequence motif:

```
defconcept motif1
     (image 5ADH-hexamer-203
     [distance,delta]);
```

asserting that this symbolic image preserves the topological distance and Δ relations. 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 operators

Generalization is based on three nondeterministic structural transformations applied to image terms. For any symbolic image I and relation set R, the image term (I, R) can be generalized in various ways:

Rule 1: by replacing a part p of a component (p, c) in I by a more general concept term q, that is,

$$\frac{q \succeq_T p}{(\mathtt{replace}(I,(p,c),(q,c)),R) \succeq_T (I,R)},$$

Rule 2: by deleting one or more parts from the image I, that is,

$$\frac{}{(\mathtt{delete}(I,C),R) \succeq_T (I,R)},$$

Rule 3: by replacing a relation r in R by a more general relation r', that is,

$$\frac{r' \succeq_T r \qquad r \in R}{(I,(R \cup \{r'\}) - \{r\}) \succeq_T (I,R)},$$

Rule 4: by removing relation identifiers from the relation set R, that is,

$$\frac{R' \subseteq R}{(I,R') \succeq_T (I,R)}.$$

For example, a motif more general than motif1 (defined above) can be constructed by a single application of Rule 1:

```
defconcept motif2 (image
   (replace 5ADHhexamer203
       [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) $\succeq_T$ Glycine). The extension of motif2 includes not only instances with a glycine in a relative position; any residue that is both small and hydrophobic (e.g., alanine, threonine) can be substituted in that position. Thus motif2 $\succeq_T$ motif1, and this is depicted by the concept taxonomy of Figure 2.

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.

The IMEM (Image MEMory) system (Conklin and Glasgow, 1992) is a similarity–based structured concept formation system which discovers, revises, maintains and organizes an $\mathcal{SDL}$ knowledge base of images. The system has been used to discover and describe the various configurations of hexopyranose molecules (Conklin et al., 1992). It has also been applied to an initial small set ($\sim$ 100) of fragments from the Protein Data Bank. The resulting $\mathcal{SDL}$ concept terms are presently stored in frame data structures. We estimate that tens of thousands of individual and concept frames will be necessary for a complete indexing of the Protein Data Bank. Supportive knowledge base management tools are currently under development.

Discussion

This paper has presented a representation for protein motifs which captures sequence and structure motifs in a common format, and interprets them using a uniform semantics. In addition to the syntax and semantics of our protein motif representation, their pragmatics also have to be considered. Protein sequence motifs can facilitate the incremental acquisition of sequence data into knowledge bases organized according to sequence similarity (Taylor, 1986). Protein structure motifs can be used as building blocks for protein model building in crystallography (Jones and Thirup, 1986; Claessens et al., 1989). Finally, protein structure-sequence motifs can be used for structure prediction, model building, and protein design (Unger et al., 1989).

Our research in the area of protein motif discovery is progressing in conjunction with a project in *molecular scene analysis* (Fortier et al., 1993), which is concerned with the automated reconstruction and interpretation of crystal and molecular structures. A key problem-solving approach in the framework is *reasoning by analogy*, where existing molecular fragments are used to anticipate, predict, and evaluate partial interpretations of molecular scenes.

Information retrieval theory has shown that a hierarchical clustering can improve both the precision and recall of data objects (Salton and Wong, 1978). Similarity searches on clustered files can be conducted rapidly because large numbers of instances are rejected by the indexing structure. In our scheme, the cluster hierarchy is a subsumption taxonomy where concept names are associated with image terms. Similar fragments are retrieved by *classifying* the query fragment, and returning other fragments with the same classification.

A discovered concept taxonomy of protein structure-sequence motifs will be used for two main purposes in our molecular scene analysis framework. First, it represents a library of consensus sequence motifs that may occur in new proteins. Queries will supply sequence information, and the system will propose structure associations. Second, a taxonomy of concepts can be viewed as a library of common structure motifs. Queries in this case will supply a partially interpreted map and hypothesized fragment; the system will provide analogous structure motifs, and predict sequence motifs corresponding to that structure. It is also possible to rephrase the classification problem slightly, obtaining a mechanism for *model-driven segmentation* of the map. Initial studies (Conklin et al., 1993) used a taxonomy of small molecular motifs to guide the interpretation of a small high-resolution electron density map, and showed the potential of this technique. During the course of structure determination, all classification query types will be performed many times, thus it is essential that efficient indexing methods be used.

In summary, a formal representation and rigorous semantics for protein motifs is important for a number of problems in molecular biology, and crucial to our approach to molecular scene analysis. This paper has surveyed other representations, found them to be lacking in semantic uniformity, and has presented a new technique based on description logics. This research is a step towards an integrated framework in which all types of protein motifs and their abstractions can be expressed in large knowledge bases.

Acknowledgements

This research has been supported by a Postgraduate Scholarship and operating grants from the Natural Science and Engineering Research Council of Canada.

References

Bernstein, F. C.; Koetzle, T. F.; Williams, J. B.; Meyer Jr., E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 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.

Chang, C. C. and Lee, S. Y. 1991. Retrieval of similar pictures on pictorial databases. *Pattern Recognition* 24(7):675–680.

Claessens, M.; Van Cutsem, E.; Lasters, I.; and Wodak, S. 1989. Modelling the polypeptide backbone with spare parts from known protein structures. *Protein Engineering* 2(5):335–345.

Cohen, F. E.; Ararbanel, R. M.; Kuntz, I. D.; and Fletterick, R. J. 1986. Turn prediction in proteins using a pattern matching approach. *Biochemistry* 25:266–275.

Conklin, D.; Fortier, S.; Glasgow, J.; and Allen, F. 1992. Discovery of spatial concepts in crystallographic databases. In Zytkow, J. M., editor 1992, *Proceedings of the ML92 Workshop on Machine Discovery*, Aberdeen, Scotland. 111–116.

Conklin, D.; Fortier, S.; and Glasgow, J. 1993. Knowledge discovery in molecular databases. *IEEE Transactions on Knowledge and Data Engineering: Special Issue on Learning and Discovery in Knowledge-Based Databases.* To appear.

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

Fortier, S.; Castleden, I.; Glasgow, J.; Conklin, D.; Walmsley, C.; Leherte, L.; and Allen, F. 1993. Molecular scene analysis: The integration of direct-methods and artificial-intelligence strategies for solving protein crystal structures. *Acta Crystallographica* D49:168–178.

Glasgow, J. I. and Papadias, D. 1992. Computational imagery. *Cognitive Science* 16(3):355–394.

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

Harris, N.; Hunter, L.; and States, D. 1992. Megaclassification: Discovering motifs in massive datastreams. In *Proc. AAAI-92.* AAAI/MIT Press. 837–842.

Hunter, L. and States, D. J. 1991. Bayesian classificaion of protein structural elements. In Hunter, L., editor 1991, *Proc. Seventh IEEE Conf. on AI Applications: The Biotechnology Computing Minitrack.*

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

Lathrop, R. H.; Webster, T. A.; and Smith, T. F. 1987. ARIADNE: Pattern-directed inference and hierarchical abstraction in protein structure recognition. *Communications of the ACM* 30(11):909–921.

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

Nebel, B. 1990. *Reasoning and revision in hybrid representation systems.* Springer-Verlag.

Rooman, M. J.; Wodak, S. J.; and 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.; and Wodak, S. J. 1990a. Automatic definition of recurrent local structure motifs in proteins. *J. Mol. Biol.* 213:327–336.

Rooman, M. J.; Rodriguez, J.; and Wodak, S. J. 1990b. Relations between protein sequence and structure and their significance. *J. Mol. Biol.* 213:337–350.

Sali, A. and Blundell, T. 1990. Definition of general topological equivalence in protein structures. *J. Mol. Biol.* 212:403–428.

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

Smith, R. F. and Smith, T. F. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci.* 87:118–122.

Sternberg, M. J. E. and Islam, S. A. 1990. Local protein sequence similarity does not imply a structural relationship. *Protein Engineering* 4(2):125–131.

Taylor, W. R. 1986. Consensus template alignment. *J. Mol. Biol.* 188:233–258.

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

Unger, R.; Harel, D.; Wherland, S.; and 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 1990, *Concepts and applications of molecular similarity.* John Wiley & Sons. chapter 3.