# Knowledge Discovery of Multilevel 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**

A new category of protein motif is introduced. This type of motif captures, in addition to global structure, the nested structure of its component parts. A dataset of four proteins is represented using this scheme. A structured machine discovery procedure is used to discover recurrent amino acid motifs and this knowledge is utilized for the expression of subsequent protein motif discoveries. Examples of discovered multilevel motifs are presented.

## Introduction

In an earlier paper (Conklin, Fortier, & Glasgow 1993) we presented a knowledge representation formalism for protein motif representation and discovery called a *spatial description logic* ($\mathcal{SDL}$). The logic is specifically tailored to reasoning about spatially structured objects, and is a convenient and expressive formalism for representing various types of protein motifs. Our knowledge discovery method, a relational conceptual clustering procedure, uses $\mathcal{SDL}$ as a representation language and groups similar protein fragments into classes described by discovered, subsuming motifs. This representation and discovery technology has been applied to several molecular datasets.

The spatial description logic formalism is used, in general, to describe *structured objects*, that is, objects with parts along with defined relations among these parts. Parts can recursively refer to other structured objects, providing a mechanism for nesting or reducing the complexity of an object. In our previous research, nested motifs were not explored. The purpose of this paper is to illustrate the use of $\mathcal{SDL}$ on *multilevel* protein motif representation and discovery. In a multilevel representation language, the parts of a protein motif (amino acids) need not be primitive identifiers, rather, they can be structured objects - - amino acid motifs — in turn. Both protein motifs and embedded amino acid motifs are discovered by our machine learning procedure. The advantage of the multilevel representation is that the structure at one level provides contextual information that has some bearing on structure at a lower or higher level.

The first section of this paper reviews protein motif representation using $\mathcal{SDL}$, and presents the concept of a multilevel structured protein motif. The second section reviews the machine discovery procedure, and in the third section it is applied to a small database of multilevel protein fragments. Examples of discovered multilevel protein motifs are presented. The paper concludes with a discussion of the pragmatics of multilevel protein motifs, and potential future research.

## Multilevel protein motif representation

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 structural elements 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. Finally, *structure-sequence* motifs are combined 1d-3d structures that associate sequence information with a structure motif. Structure-sequence motifs need not indicate a fixed direction of implication between structure and sequence. This is particularly useful for our purposes (Fortier *et al.* 1993), since they represent, in addition to sequence-structure rules, 3d features that may be matched to fragments in an emerging protein electron density map.

There has been much research on the discovery of pure structure motifs, that is, motifs with no attached sequence information [e.g., (Hunter & States 1991; Rooman, Rodriguez, & Wodak 1990; Onizuka *et al.* 1993)]. Recent research has looked at structure-sequence motifs, which are also concerned with the *characteristics* of a residue at a particular position of a structure motif [e.g., (Conklin, Fortier, & Glasgow 1993; Unger *et al.* 1989; Zhang *et al.* 1993)].

All previous structure-sequence discovery work, including our own, 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). The approach we have chosen is to use a machine discovery procedure to autonomously discover its own rotamer classes. Both approaches require a knowledge representation, such as $\mathcal{SDL}$, capable of describing multilevel structured objects.

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 otherwise atomic (not further decomposable). 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. To draw an example from protein structure, atoms are level-0 objects. Amino acids are level-1 objects, containing only level-0 objects as parts. Protein structure motifs — polypeptide chains of amino acids — are either level-1 or level-2 objects, depending on whether amino acids are atomic or composite in the representation.

Figure 1 illustrates the two styles of structure-sequence protein motif. The level-1 protein motif in Figure 1 (top) [see (Conklin, Fortier, & Glasgow 1993) for a full discussion of level-1 motifs] has as parts primitive concept terms such as `arginine` and `polar` and `hydrophobic` These concepts are not themselves 3d motifs and have no internal structure. Each amino acid is positioned at its C$\alpha$ location. The motif is a 4d object; dimensions 1 through 3 are used for Cartesian coordinates, and dimension 4 for the sequence position. The motif preserves two relations, the binary topological `distance` relation and the quaternary spatial `delta` relation. The `distance` relation measures how far apart two residues are in the sequence. The `delta` relation partitions the virtual backbone torsion angle space into four ranges, defined by Ring *et al.* (1992): `U` (-75 to 15), `L` (15 to 105), `Z` (105 to 195) and `J` (195 to 285). To the right of the motif is the $\mathcal{SDL}$ syntax for generating the depiction and relational semantics. The keyword `image` in the concept definition declares an image term: this is a method for concisely expressing structured concepts and relations.

Figure 1 (middle) illustrates a level-2 motif: the type we are concerned with in this paper. One of the parts of the motif (`aamotif-1`) is itself an amino acid motif with an internal level-1 structure. The amino acid motif is illustrated below the container motif. It preserves the `bonded` and `planarity` relations, and subsumes any phenylalanine with the indicated parts in the appropriate topological and planarity relationships. The `planarity` relation (Klyne & Prelog 1960) divides torsion angle space into four regions: `syn-periplanar` (-30 to +30), `anti-periplanar` (150 to 210), `+clinal` (30 to 150) and `-clinal` (210 to 330). Note that the level-2 motif of Figure 1 is only one of many possible ways of parsing the level-0 atomic structure into aggregates. It is, however, quite natural since amino acids are the accepted building blocks of proteins. While the internal relations of the amino acids — the intramolecular relations — are preserved by a level-2 motif, "cousin" relationships — interatomic relations between different amino acids of the container motif — are not retained. This is the small penalty that must be paid with aggregation.

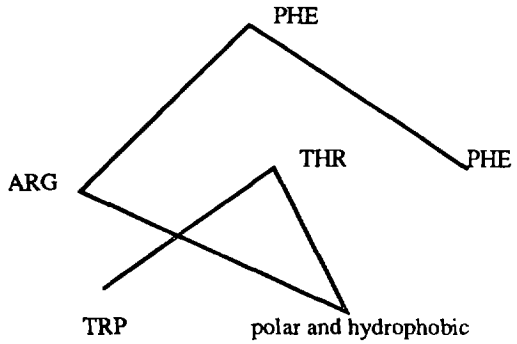
Not illustrated by Figure 1 is a protein motif represented as a level-0 object. Although this is possible in $\mathcal{SDL}$, there are many problems with such a representation. Objects would be complex and hard to understand due to the lack of conceptual organization or grouping of parts. Furthermore, it becomes computationally difficult to match objects with many atomic parts. Researchers in chemical information systems have also encountered similar problems, and have considered "reduced" graphs as a representation language for small molecules (Takahashi, Sukekawa, & Sasaki 1992). However, in contrast to the $\mathcal{SDL}$ language, reduced graphs discard the internal structure of the aggregated parts.

A central idea in $\mathcal{SDL}$, and indeed in all description logics, is *subsumption*. One concept is subsumed by another if all of its possible instances are also instances of the other. The concept definitions and the semantics of a particular description logic dictate the criteria for instance relationships. As outlined in (Conklin, Fortier, & Glasgow 1993), subsumption in $\mathcal{SDL}$ can be computed by finding a relational monomorphism (Haralick & Shapiro 1993) — similar to a subgraph isomorphism but with hyperedges — which also preserves subsumption among the parts of a motif. Since the parts of a motif can be any concept, including another motif, this inductive definition of subsumption extends immediately and elegantly to multilevel protein motifs.

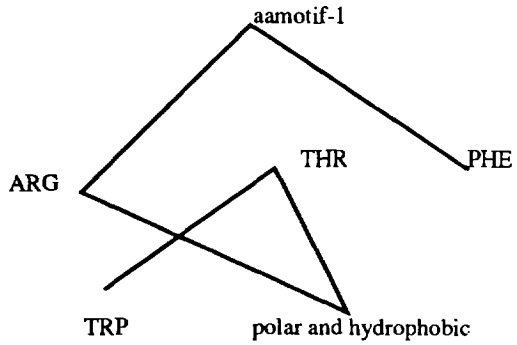
## Multilevel protein motif discovery

In an earlier paper (Conklin & Glasgow 1992) we described IMEM (Image MEMory), an incremental relational conceptual clustering system. IMEM is a prototype, similarity-based discovery system; observed structural similarities are assumed to be potentially indicative of an interesting and useful discovery. The system scans $\mathcal{SDL}$ concept definitions, one by one, and incorporates them into an expanding concept taxonomy. This concept taxonomy is used to direct a motif towards similar motifs, when concept formation may be triggered by high similarity. The newly formed motif, a common subsumer, is then *classified*: placed just below all most specific subsumers and just above all most general subsumees. In this manner a network of recurrent motifs emerges and is maintained.

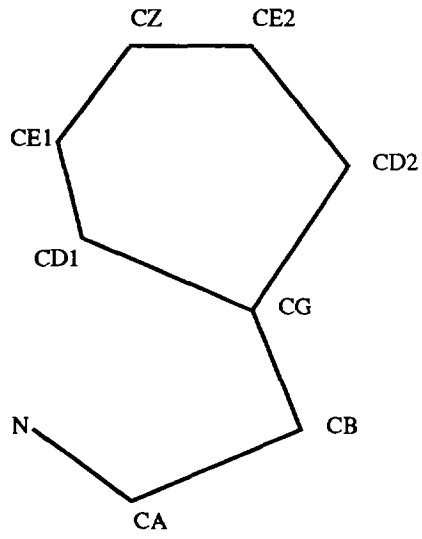
The IMEM conceptual clustering system has been




```
defconcept  level-1-motif (image (
        [TRP,[5.8,5.4,14.6,37]]
        [THR,[7.5,1.9,15.0,38]]
        [polar and hydrophobic,[8.7,3.6,18.4,39]]
        [ARG,[5.2,2.9,19.8,40]]
        [PHE,[6.1,-0.7,21.2,41]]
        [PHE,[9.4,0.5,22.6,42]])
  [distance,delta]);
```




```
defconcept  level-2-motif (image (
        [TRP,[5.8,5.4,14.6,37]]
        [THR,[7.5,1.9,15.0,38]]
        [polar and hydrophobic,[8.7,3.6,18.4,39]]
        [ARG,[5.2,2.9,19.8,40]]
        [aamotif-1,[6.1,-0.7,21.2,41]]
        [PHE,[9.4,0.5,22.6,42]])
  [distance,delta]);
```




```
defconcept  aamotif-1 (image (
        [N,[8.1,0.4,22.2]]
        ...
        [CZ,[10.4,-2.9,18.8]])
  [bonded,planar]) and PHE;
```

**Figure 1:** Two styles of protein motif. Top: a level-1 motif. Middle: a (multi) level-2 motif. Bottom: one of its components.

applied with success to several small molecule datasets from the Cambridge Structural Database (Allen *et al.* 1991), and produces results which compare very favourably with human and/or other classification schemes (Conklin *et al.* 1992). IMEM has also been applied to a medium-sized dataset of level-1 protein fragments (results forthcoming).

The $\mathcal{SDL}$ interpreter uses a reference-by-meaning semantics. The meaning of an identifier (i.e., a concept name) is fixed at definition time, and an identifier must be defined — occurring on the left hand side of a definition — before being used in another concept definition. Therefore, for multilevel protein motifs, all amino acids must be defined before any level-2 container motif is encountered.

Central to this discovery method is the computation of a common subsumer of two structured objects which have high similarity. Since similarity is computed by finding a common subimage, it requires a structure-preserving mapping between the parts of the two images. Similarity of multilevel objects is measured, in the current system, by inspecting only the relational structure and not the common characteristics of parts. Another issue that arises, for multilevel motifs, is that of generalization or generating a common subsumer for corresponding level-1 parts. One choice is to generate a new subsuming level-1 motif whenever necessary. This can make the learning process tedious since the common subsumer routine is computationally expensive, and also many duplicate concepts could be created. The more natural choice, consistent with the generalization method for parts which are not motifs, is to use the current concept taxonomy to return a more general concept term. This works for multilevel motifs because all level-1 structures will have been clustered (new subsuming concepts created) before any level-2 motif is encountered. Generalization of two parts relative to the current concept taxonomy is done by finding the conjunction of all their least upper bounds. The relational conceptual clustering system of Thompson & Langley (1991) uses a similar generalization method, although their concept taxonomy is a tree, and not a more general partial order as in $\mathcal{SDL}$.

## Results

A database of 402 overlapping protein heptamer fragments was created from four proteins [Protein Data Bank codes 4HHB (chain B, 140 fragments), 5PTI (52 fragments), 1BP2 (117 fragments) and 1PCY (93 fragments)]. Taylor's (1986) domain theory of amino acid physicochemical properties was coded as background knowledge in $\mathcal{SDL}$. The 402 heptamer fragments contained a total of 426 amino acids; these were also extracted from the PDB with their complete atomic 3d coordinates. The names assigned to the atoms are the PDB labels (given in PDB fields 14-16 of ATOM records); this predefined labelling greatly simplifies the relational matching process for amino acids.

| Parts | | | | Planarity relation |
|---|---|---|---|---|
| CE2 | CD2 | CG | CB | ap |
| CE1 | CD1 | CG | CB | ap |
| CD2 | CG | CB | CA | -c |
| CD1 | CG | CB | CA | +c |
| CG | CB | CA | N | -c |
| CG | CB | CA | C | ap |

Table 2: The internal planarity relationships of the rightmost motif of Figure 2.

### Level-1 (amino acid) classification

All 426 amino acids were first incorporated into the initial knowledge base comprising the domain theory indicated above. We used a low threshold for concept formation, so that very slight similarities will trigger the generation of an amino acid motif. The computation of a common subsumer takes place only between amino acids of the same type. A total of 96 level-1 amino acid motifs were discovered by our system. Each amino acid type has its own sub-taxonomy, which can be stored and, in the future, incrementally refined. For example, Table 1 illustrates the concept taxonomy for proline. The first entry in a row is the name of the motif; concepts appear with a unique name, and individuals are at the leaves of the taxonomy. There are 7 discovered concepts and 21 instances in this subtaxonomy. The second entry in a row is the number of parts in the motif; note that the number of parts decreases as one climbs the taxonomy.

To illustrate subsumption of amino acid motifs, Figure 2 displays three discovered phenylalanine motifs, ordered by subsumption. The motif at the left of the figure is a planar six-member ring with an attached carbon. This motif occurs in 22 out of 23 phenylalanine amino acids encountered in the training set. The rightmost motif is very specific, having 10 parts in certain planarity relationships. The internal planarity relationships of this motif, excluding the planar ring which is all syn-periplanar, are given in Table 2.

### Level-2 (heptamer) classification

The discovered concept taxonomy of amino acids provides extra background knowledge for protein motif discovery. All 402 heptamer fragments were incorporated into this knowledge base. The generalization method outlined earlier was used.

An example of a discovered multilevel motif is given in Figure 3. One of its parts is a previously discovered, specific amino acid proline motif (recall Table 1). The motif is a strand with two J relationships (abbreviated as the structural sequence JJ). In the small database of 402 heptamers, this motif has 6 instances. This is an interesting discovery, as it illustrates that the structure and sequence of this particular motif is often associated with a particular proline rotamer.

```
UNIQ-129 3 PRO
        UNIQ-73 4 PRO
                4HHB-aa5 7 PRO
                UNIQ-143 6 PRO
                        4HHB-aa58 7 PRO
                        4HHB-aa36 7 PRO
                        4HHB-aa100 7 PRO
                        1BP2-aa14 7 PRO
                        1BP2-aa18 7 PRO
                        1BP2-aa68 7 PRO
                        1PCY-aa86 7 PRO
                5PTI-aa2 7 PRO
                1BP2-aa37 7 PRO
                1BP2-aa110 7 PRO
                1PCY-aa36 7 PRO
        UNIQ-369 4 PRO
                4HHB-aa51 7 PRO
                UNIQ-374 5 PRO
                        UNIQ-1380 6 PRO
                                1PCY-aa23 7 PRO
                                4HHB-aa125 7 PRO
                        UNIQ-1506 6 PRO
                                1PCY-aa47 7 PRO
                                4HHB-aa124 7 PRO
                5PTI-aa8 7 PRO
                5PTI-aa9 7 PRO
                5PTI-aa13 7 PRO
                1PCY-aa16 7 PRO
```

Table 1: The discovered sub-taxonomy of proline motifs.

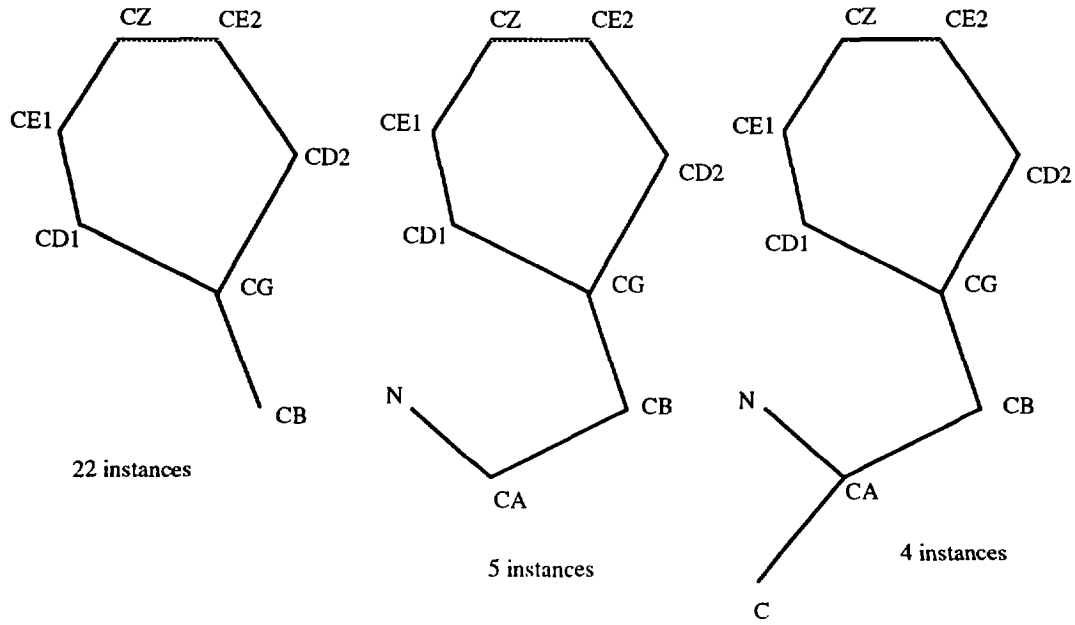



Figure 2: Examples of discovered level-1 (PHE) motifs.

| Id | Sequence | Structure | (M+,M-) | DRMS |
|---|---|---|---|---|
| 1819 | (X)(X)(UNIQ-4)(l)(UNIQ-9)(h)(X) | LLLL | (2,0) | 0.40 |
| 1828 | (s)(pay)(UNIQ-143)(h)(l)(c)(h) | JLLL | (2,0) | 0.64 |
| 1954 | (X)(UNIQ-843)(UNIQ-138)(h)(c)(h)(h) | JZLL | (2,0) | 0.73 |
| 2008 | (pd)(s)(nay)(UNIQ-143)(X)(X)(X) | JJLL | (2,0) | 0.90 |
| 2017 | (h)(UNIQ-4)(h)(s)(spay)(UNIQ-143)(pa) | JZJJ | (2,0) | 0.33 |
| 2026 | (hs)(h)(UNIQ-39)(h)(h)(s)(c) | LLLL | (2,0) | 0.87 |
| 2035 | (X)(UNIQ-4)(h)(s)(s)(l)(l) | LLLL | (2,0) | 0.19 |
| 2233 | (X)(h)(psda)(UNIQ-129)(X)(UNIQ-150)(pd) | ZJLL | (2,0) | 0.81 |
| 2386 | (pay)(s)(X)(X)(s)(UNIQ-113)(X) | LLLL | (3,1) | 0.48 |
| 2494 | (l)(H)(l)(psa)(UNIQ-73)(E)(pay) | ZJJL | (2,0) | 0.88 |
| 2854 | (X)(X)(h)(h)(X)(X)(UNIQ-4) | LLLL | (10,2) | 0.43 |

Table 3: Some discovered multilevel structure-sequence motifs.

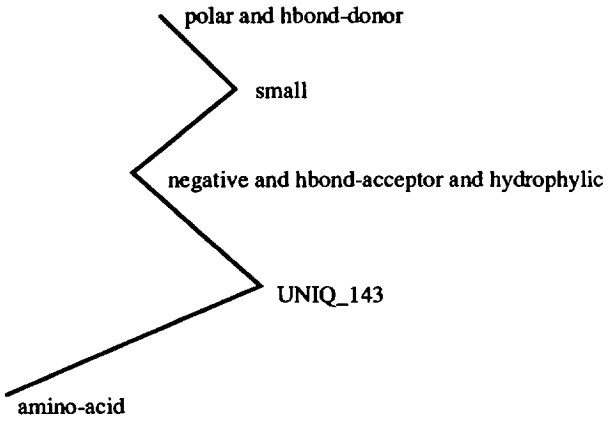



Figure 3: An example of a discovered level-2 motif.

Table 3 displays more discovered motifs in a textual notation. The "sequence" column shows seven elements enclosed by brackets; each element is a conjunction of property abbreviations, e.g., a "psa" residue is polar, small, and a hydrogen bond acceptor. X indicates any residue: other capital letters are abbreviations for specific residues. The "structure" column shows the structural sequence of the motif. Although many motifs were discovered, Table 3 displays only those with certain features. First, they must contain a discovered amino acid motif as a component. Second, the average similarity of their instances, as measured by a distance RMS metric, must be less than 1.0 Angstrom. This is done because it does not necessarily follow that motifs, qualitatively similar according to the `delta` relation, are also quantitatively or "visually" similar. We also quantify the relationship between sequence and structure of the motif. The sequence of motif 2854, for example, occurs 12 times in the small database; 10 times (M+) in the indicated structure LLLL, and 2 times (M-) in some other structure. Clearly this is an interesting motif and discovery, as there is some confidence and support for a sequence–structure prediction rule. The motif UNIQ-4 is a LEU motif. The component UNIQ-113 of motif 2386 is a PHE and is the leftmost motif displayed in Figure 2. Also interesting is motif 2017, which is subsumed by the motif of Figure 3.

## Discussion

This paper has described and applied a representation and discovery system for finding spatial regularities among objects. Protein motifs are represented as multilevel structured objects, where components can have an internal structure. This allows a discovery procedure to capture associations between the spatial structure of a motif, its sequence, and the nested spatial structure of its parts.

The discovery system can produce motifs for which there is a near exclusive relationship between sequence and structure. These motifs might be used for structure prediction or could be matched directly to an emerging electron density map. Multilevel motifs add an extra dimension to this analysis. The parts of a multilevel motif can be focused on, and they may tell us where to look for atomic parts in a higher resolution map.

An issue that has not yet been totally resolved is how exactly to quantify the "interestingness" of a discovery. Many motifs, varying in specificity, may be produced by our discovery procedure. In this paper, "good" amino acid motifs are ones that are specific and recurrent in the database. Similarly, good multilevel motifs are specific, and their sequences have a preferential relationship with an associated structure. Certainly, specific multilevel motifs with many instances represent interesting patterns, while overly general motifs might not. However, the value of a particular motif will often depend on the eventual use of the knowledge base in general. For motif retrieval purposes, even very general motifs can play an important role in the index-

ing of protein fragments.

Finally, although we have only considered level-2 protein fragments — segments of level-1 amino acid residues — the scheme can also be applied to higher level structure. For example, it would be interesting to see if certain supersecondary (level-3) structures have preferential residue segment (level-2) motifs. For such an exercise, it might be necessary to extend $\mathcal{SDL}$ to represent line and volume data, and not only point data objects.

## Acknowledgements

DC would like to thank the Cambridge Crystallographic Data Centre for providing the computing facilities on which much of this work was carried out. This research has been supported by a Postgraduate Scholarship and operating grants from the Natural Science and Engineering Research Council of Canada.

## References

Allen, F. H.; Davies, J.; Galloy, J.; Johnson, O.; Kennard, O.; Macrae, C.; Mitchell, E.; Mitchell, G.; Smith, J.; and Watson, D. 1991. The development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comp. Sci.* 31:187–204.

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

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

Conklin, D.; Fortier, S.; and Glasgow, J. 1993. Representation for discovery of protein motifs. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 101–108. AAAI/MIT Press.

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.

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

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

Klyne, W., and Prelog, V. 1960. Description of steric relationships across single bonds. *Experientia* 16:521–523.

Onizuka, K.; Ishikawa, M.; Wong, S. T. C.; and Asai, K. 1993. A multi-level description scheme of protein conformation. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 301–309. AAAI/MIT Press.

Ponder, J. W., and Richards, F. M. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775–791.

Ring, C. S.; Kneller, D. G.; Langridge, R.; and Cohen, F. E. 1992. Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* 224:685–699.

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

Takahashi, Y.; Sukekawa, M.; and Sasaki, S. 1992. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *Journal of Chemical Information and Computer Sciences* 32:639–643.

Taylor, W. R. 1986. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* 188:233–258.

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

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.

Zhang, X.; Fetrow, J. S.; Rennie, W. A.; Waltz, D. L.; and Berg, G. 1993. Automated derivation of substructures yields novel structural building blocks in globular proteins. In Hunter, L.; Searls, D.; and Shavlik, J., eds., *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 438–446. AAAI/MIT Press.