Motif Discovery in Protein Structure Databases[1]

Janice Glasgow, Evan Steeg and Suzanne Fortier
Queen's University

Abstract

This chapter overviews the topic of protein motif discovery. It presents current approaches to knowledge discovery, focusing on their applications to the protein domain. In general, a motif is considered an abstraction over a set of recurring patterns in a dataset. Although we are primarily concerned with protein structure motifs, the chapter also considers sequence motifs and combinations of sequence/structure motifs. The research described is motivated by our need to organize and understand the rapidly growing protein databases. Discovered motifs are also useful in automating the process of structure determination from crystallographic databases.

The field of **knowledge discovery** is concerned with the theory and processes involved in the representation and extraction of patterns or motifs from large databases. Discovered patterns can be used to group data into meaningful classes, to summarize data or to reveal deviant entries. Motifs stored in a database can be brought to bear on difficult instances of structure prediction or determination from X-ray crystallography or NMR experiments. The need for automated discovery techniques is central to the understanding and analysis of the rapidly expanding repositories of protein sequence and structure data.

This chapter deals with the discovery of protein structure motifs. A **motif** is an abstraction over a set of recurring patterns observed in a dataset; it captures the essential features shared by a set of similar or related objects. In many domains, such as computer vision and speech recognition, there exist special regularities that permit such motif abstraction. In the protein science domain, the regularities derive from evolutionary and biophysical constraints on amino acid sequences and structures. The identification of a known pattern in a new protein sequence or structure permits the immediate retrieval and application of knowledge obtained from the analysis of other proteins. The discovery and manipulation of motifs — in DNA, RNA, and protein sequences and structures — is thus an important component of computational molecular biology and genome informatics. In particular, identifying protein structure classifications at varying levels of abstraction allows us to organize and increase our understanding of the rapidly growing protein structure datasets. Discovered motifs are also useful for improving the efficiency and effectiveness of X-ray crystallographic study of proteins, for drug design, for understanding protein evolution, and ultimately for predicting the structure of proteins from sequence data.

Motifs may be designed by hand, based on expert knowledge. For example, the Chou-Fasman protein secondary structure prediction program (1978), which dominated the field for many years, depended on the recognition of pre-defined, user-encoded sequence motifs for α-helices and β-sheets. Several hundred sequence motifs have been catalogued in Prosite (Bairoch 1991); the identification of one of these motifs in a novel protein often allows for

[1] From Glasgow, J.I., Steeg, E. and Fortier, S., "Motif Discovery in Protein Structure Database," in "Pattern Discovery in Molecular Biology: Tools, Techniques and Applications," by Jason Wang, Bruce Shapiro and Dennis Sha sha (Eds.). Reprinted by permission of Oxford University Press, Inc.

immediate function interpretation. In recent years there has been much interest and research in automated motif discovery. Such work builds on ideas from machine learning, artificial neural networks, and statistical modeling, and forms the basis for the methods for protein motif discovery described in this paper.

The search for protein structure motifs begins with the knowledge that some proteins with low sequence similarity fold into remarkably similar 3D conformations. Moreover, even globally different structures may share similar or identical substructures (Rooman, Rodriguez, & Wodak 1990; Unger *et al.* 1989), as predicted by Pauling, Corey & Branson h (1951) and verified in many of the early crystallographic experiments. Researchers have discovered and catalogued important motifs at the **secondary**, **super-secondary**, and **tertiary structure** levels (Ponder & Richards 1987; Taylor & Thornton 1984; Wilmot & Thornton 1988). Recently, there has also been interest in the automated discovery and use of structural motifs at levels finer than that of secondary structure (Hunter & States 1991; Rooman, Rodriguez, & Wodak 1990; Unger *et al.* 1989).

In the remainder of this chapter, we describe several types of protein motifs and present approaches to knowledge discovery and their application to the problem of structure motif discovery. We conclude the chapter with a presentation of criteria by which a motif, and associated motif discovery techniques, may be judged.

Protein Motifs

The study 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 over one or more fragments drawn from proteins of known sequence and tertiary structure. The Protein Data Bank (Bernstein *et al.* 1977) is the main source of data regarding the 3D structure of proteins.

Protein motifs can be roughly divided into four categories (as illustrated in Figure 1): **sequence motifs** are linear strings of **amino acid residues** (or residue classes) with an implicit topological ordering; **sequence-structure** motifs are sequence motifs that associate residues in the motif with secondary structure identifications; **structure motifs** are 3D objects that correspond to portions of a protein backbone, possibly combined with side-chains; and **structure-sequence motifs** are structure motifs in which nodes of the graph are annotated with sequence information. The distinction between sequence and structure motifs has previously been considered in (Thornton & Gardner 1989) and (Conklin, Fortier, & Glasgow 1993).

Different types of motifs have different purposes. For example, 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 & Thirup 1986). Finally, protein structure-sequence motifs are useful in automated structure prediction, model building and model design (Unger *et al.* 1989). They are also applicable to automated approaches to protein structure determination from crystallographic data (Fortier *et al.* 1993; Glasgow, Fortier, & Allen 1993).

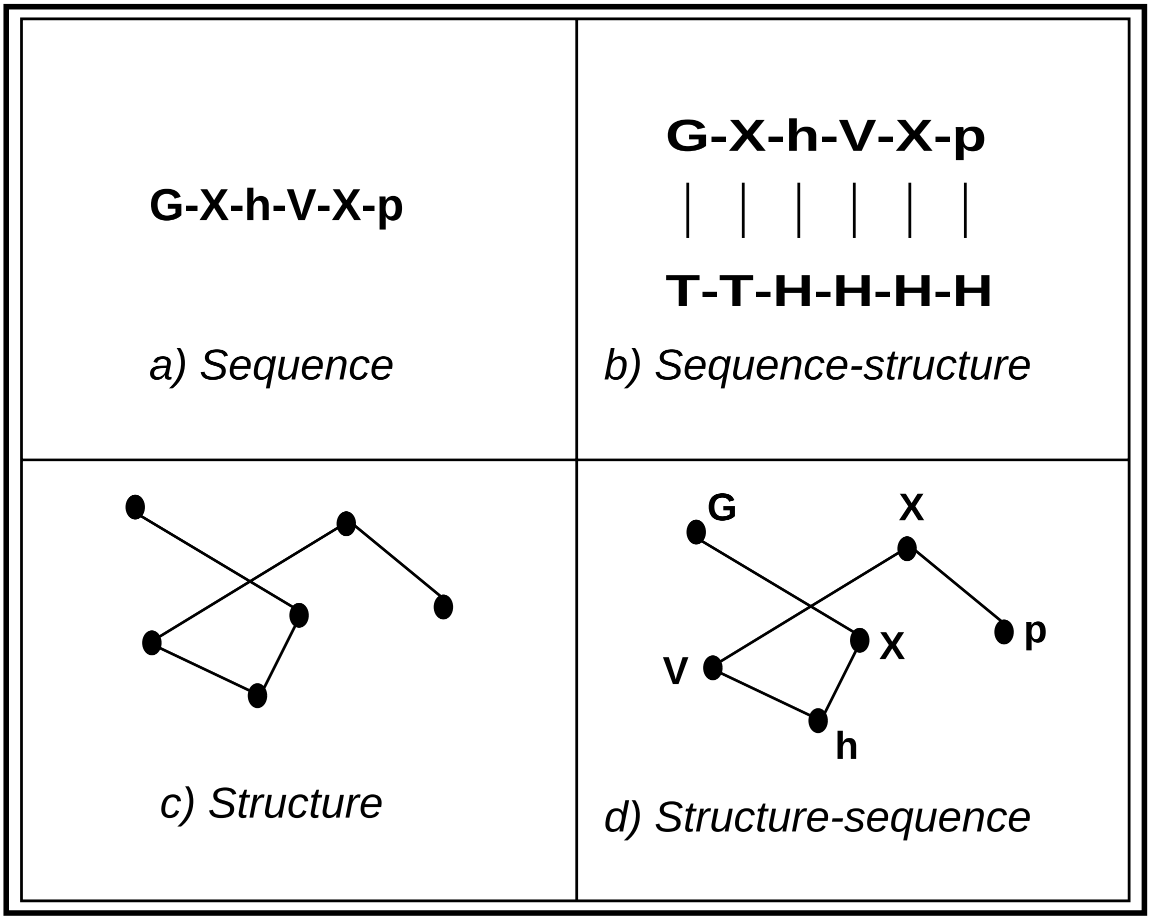

Figure 1: Various types of protein motifs. Legend: **X**: any residue, **G**: glycine, **V**: valine, **H**: α-helix,**T**: turn, **p**: polar, **h:** hydrophobic.

Following, we discuss the different types of protein motifs in more detail and present recent studies that involved the discovery of such motifs.

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 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).

Finding motifs in sequences involves two main steps: 1) assembling a training set of sequences with common structure or function, then 2) analyzing the training set for regions of conserved amino acids. Sequence motifs may be discovered from a maximal alignment of one or more protein sequences, followed by the abstraction of residues at aligned positions. Conserved residues are those identical at corresponding alignment positions. In recent years, the use of **Hidden Markov Models** has become predominant in the sequence modeling area (Krogh *et al.* 1994). There is an extensive literature on the comparison of sequence motifs.

Much of the work on protein 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 (1989) as "structure-related sequence motifs") have an inherent directionality of implication from sequence to structure. The work of Rooman and Wodak (1988) associates each amino acid in a motif with a standard secondary structure identifier (e.g., motif b in Figure 1). This work demonstrated that, while associations sufficient to predict a complete protein structure were not derived, a number of reliably predictive motifs do exist.

Structure and structure-sequence motifs

Structure motifs constitute the building blocks that can be used to describe the tertiary structure of a protein. The accurate prediction of protein tertiary structure from amino acid sequence, while theoretically possible, remains one of the great open problems in molecular biology. One way of addressing this has been to break the problem into the problem of predicting secondary structure from sequence followed by the problem of packing the secondary structure into 3D conformations.

Although secondary structures are the most commonly considered structure motifs, there has been increased interest in the automated discovery and use of structural patterns at both a finer and coarser level than that of α-helix or β-sheet. Unger et al. (1989), for example, report the discovery of motifs for structural fragments containing 6 amino acids, whereas Taylor and Thornton (1984) consider motifs at the level of super-secondary structure.

Global tertiary structure motifs typically characterize protein families and super-families (Orengo, Jones, & Thornton 1994), though several global motifs (like the so-called TIM-barrel) are shared by proteins that are not evolutionarily related nor functionally similar

(Jones & Thornton 1994).

The abundance of sequence information produced by large-scale sequencing efforts, combined with the increasing rate of structure information coming from X-ray crystallography and NMR, is producing the kind of huge and diverse databases necessary for both model-based and "memory-based" structure prediction and determination, as well as for phylogenetic analysis in evolutionary biology. Concomitant with the increasing availability of these different data types, is a growing belief that protein structure prediction and recognition methods ought to be, like the protein folding process itself, a simultaneously global and local, bottom-up and top-down application of constraints. For these reasons, the last several years have seen an increasing focus on integrated, multiple-resolution, multiple-view protein databases. This general trend and the associated opportunities to use genomic, structural, functional, and evolutionary information to reinforce each other and fill in each others' gaps have led researchers to look for structural motifs that can be associated with and predicted from sequence motifs.

A few researchers have added sequence annotation, or computed sequence-structure correlation, after completion of their structure motif discovery. Unger *et al.* (1989) used a k-nearest-neighbours method to find clusters in the space of protein backbone fragments of length 6 residues. They tabulated the frequencies of amino acids types at every position, producing a sequence motif for each of their approximately 100 structure motifs. Their results indicated that the local 3D structure of a fragment can sometimes be predicted by the assignment of the fragment to a motif based on these frequency tables. Rooman, Rodriguz & Wodak (1990) produced a physico-chemical properties motif for each of the $4-10$ structural classes they discovered in different runs of a hierarchical clustering of fragments of length $6 \ldots 10$. Zhang and Waltz (1993) used a variant of k-means clustering on structure fragments of size 7, and then tested the χ^2 significance of the association of their 23 local structure classes with particular amino acid combinations.

In contrast, other researchers have attempted to produce structure-sequence motifs, or just highly predictable structure motifs, by discovering motifs in sequence- and structure-space simultaneously. Lapedes, Steeg & Farber (1995) simultaneously trained two neural networks, one taking local sequence fragments as input, the other taking the corresponding local tertiary structure fragments, using an objective function that maximized the correlation between the two networks' outputs. The effort described in their paper, representing preliminary steps in a larger, ongoing project, succeeded in finding novel secondary structure classes that are more predictable from amino acid sequence than the standard helix, sheet, and coil classes.

Conklin (1995) used conceptual clustering to produce mutually predictable sequence and structure motifs, and by treating both sequence and structure fragments within the same model-theoretical framework, avoided the "confused dual semantics" displayed by many other attempts to relate sequence and structure.

As an alternative to the traditional approach of predicting structure from sequence, **inverse protein folding** involves predicting sequence from structure. For example, one can use genetic algorithms or other search methods to generate and test many possible sequences to see which ones might fold into a given structural pattern. Protein threading algorithms (Lathrop & Smith 1995) are used to answer the question of whether and how each such sequence “fits” onto a structure motif characterized by particular physico-chemical environments and attributes.

Hierarchical motifs

A number of groups have worked on the discovery of motifs at two or more levels of protein organization, simultaneously or in stages.

The Zhang-Waltz (1993) work is a good example of this idea. Their motif discovery process proceeded in three stages: First, they selected a set of objectively-definable primitives that compactly carried the local structural geometry of protein backbone segments. Second, the database of local structure segments, represented in terms of these new canonical feature vectors, was clustered using a k-means method. The researchers found that 23 such novel secondary structure classes produced the best fit to the data. Third and finally, they represented longer stretches (corresponding roughly to super-secondary structure) of protein backbone, in terms of sequences of symbols corresponding to the local structure classes found in step two, and designed and trained a finite-state machine to recognize the sequences. Thus the description of the trained finite state machine represented a set of *de novo* super-secondary structure motifs discovered on top of a set of *de novo* secondary structure motifs. Building upon this work the researchers later tested other input representations and clustering parameters to arrive at a set of structural building blocks that, they claim, have generality across large and diverse sets of proteins (Zhang *et al.* 1993).

Conklin, Fortier & Glasgow (1994) introduced a category of hierarchical protein motifs that captures, in addition to global structure, the nested structure of the subcomponents of the motif. This representation was used to discover recurrent amino acid motifs, which were then used for the expression of higher-level protein motifs.

Computational Approaches to Knowledge Discovery

Knowledge discovery has been defined as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad, Piatetsky-Shapiro, & Smyth 1996). Generally, the automated discovery process is interactive and iterative. Brachman and Anand (1996) divide this process into several steps. These include: developing an understanding of the application domain; creating a target data set; data cleaning and preprocessing; finding useful features with which to represent the data; data mining to search for patterns of interest; and interpreting and consolidating discovered patterns.

Several approaches to knowledge discovery have been proposed and applied. Following, we present an overview of some of these approaches derived from research in machine learning and statistics. We also discuss their applications to the problem of protein structure and structure-sequence motif discovery.

Clustering Approaches

Clustering is a discovery task that seeks to identify a finite set of classes or clusters that describe a dataset. Appropriately derived clusters can provide predictive and explanatory power, as well as lead to a better understanding of a complex dataset. Derived clusters (or classes) may be mutually exclusive or overlapping. Clustering techniques may generally be divided into three categories: numerical, statistical and conceptual, described in more detail below.

Cutting across all three major categories are other aspects by which to distinguish clustering techniques. For example, both **agglomerative** and **divisive** clustering techniques exist. Agglomerative techniques use a starting point consisting of as many clusters as instances; the starting point consists of a single cluster in divisive techniques. Clustering techniques can also be differentiated on the basis of whether they allow for overlapping clusters, or whether they only produce disjoint partitions. A clustering technique may be **incremental** or **nonincremental** depending on whether all of the observations are available at the outset of a clustering exercise. Another important distinction to be made is between approaches that incorporate flat or hierarchical representations. Hierarchical methods have a special appeal in the protein science domain, in which both the structural organization of a protein and the evolutionary process which generated it can be described hierarchically.

Numerical Clustering

In numerical clustering, samples are viewed from a geometric perspective as a set of data points in an n-dimensional space, where n is the number of attributes used to characterize each data point. The goal of the clustering exercise is to partition the data points, grouping similar points together. Figure 2 illustrates a simple clustering of a banking dataset into three overlapping classes. Such classes could be used to predict good/bad risk groups for loans.

Distance metrics are used to measure similarity/dissimilarity among data points, while criterion functions help measure the quality of the data partition. Thus, numerical clustering techniques generally rely on quantitative attributes. In structural biochemistry, attributes such as inter-atomic and inter-residue distances, and bond and torsion angles are often considered in performing numerical clustering. The objective in most numerical clustering methods is to maximize some combination of intra-class similarity and inter-class distance.

Rooman, Rodriguez & Wodak (1990) use an agglomerative numerical clustering technique to discover structure motifs, which are represented by prototypical fragments. In this method, only $C\alpha$ positions are used in the description of the amino acid position. Similarity between fragments is measured using an RMS metric of the inter-$C\alpha$ distances. A fragment is an instance of a motif class by virtue of being within an RMS distance threshold from the prototypical motif for that class.

Unger *et al.* (1989) report an experiment in structure-sequence motif discovery. Hexamers, described by $C\alpha$ positions of residues, are clustered using a $k-$nearest neighbor algorithm. As in (Rooman, Rodriguez, & Wodak 1990), similarity between hexamers is measured according to a distance metric. In this approach, structures are first aligned using a best molecular fit routine, and absolute coordinates – rather than intra-motif distances –

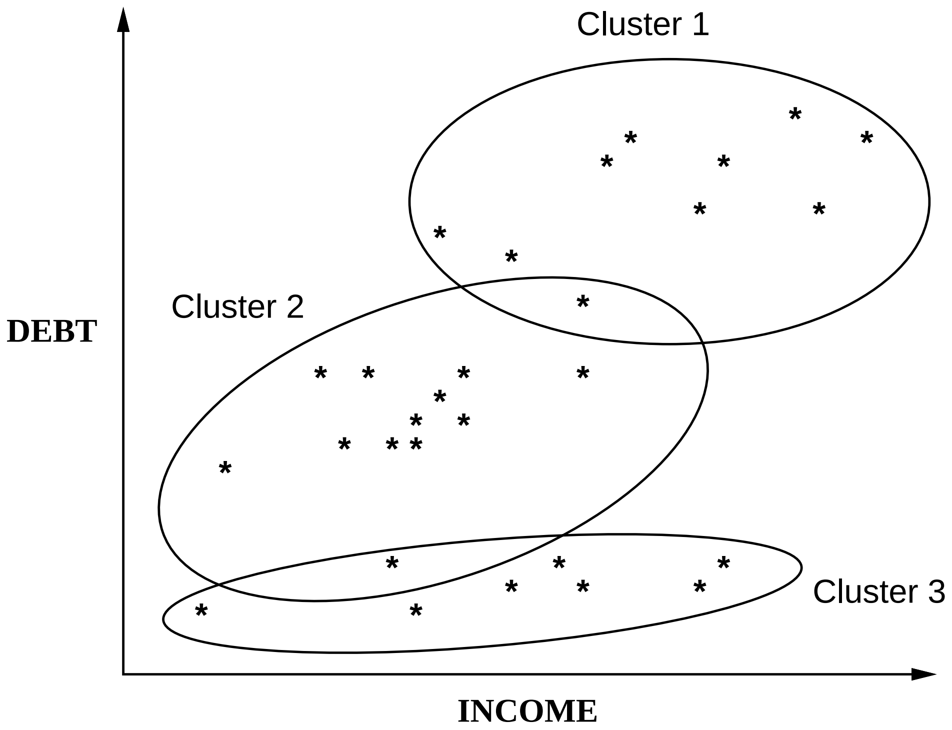

Figure 2: Clusters derived for a set of data relating income to debt to determine the viability of giving a loan.

are compared.

Bayesian Statistical Clustering

Although statistical clustering techniques typically use numerical representations and metrics, the history and philosophy behind them mandate a separate treatment. In standard numerical methods like single-linkage, total-linkage and mutual nearest neighbor clustering (Willett 1987; Jarvis & Patrick 1973), the goal is to assign data points to clusters. In contrast, the goal of a Bayesian latent class analysis, for example, is the construction of a plausible generative model of the data, that is, to explain the data.

In a latent class analysis approach to finding class structure in a set of datapoints, one begins with an underlying parameterized model. For example, one might posit that a set of points represented by a 2D scatterplot was generated by a 2D Gaussian (normal) distribution. Or the data might be better explained by a mixture, or weighted sum, of several Gaussian distributions, each with its own 2D mean vector and covariance matrix. In this approach, one tries to find an optimal set of parameter values for the representation of each datapoint. Optimality may be defined in terms of maximum likelihood, Bayes optimality, minimum description length (MDL), or minimum message length (MML). The Bayes, MDL and MML formalisms, in particular, put the somewhat vague objectives of standard numerical clustering — a tradeoff between intra-class and inter-class distances — onto firmer theoretical ground, forcing practitioners to make all assumptions and biases explicit and quantifiable. An MDL objective function, for example, imposes a cost on the total length of the shortest description of the data. This cost includes the cost of specifying the model itself, of specifying the class parameters, and of specifying the residual difference between the true positions of the datapoints versus the positions predicted from class parameters. Roughly, the model and class parameter costs can be termed complexity, and the residual differences cost can be called distortion (Buhmann & Kuhnel 1993). The complexity/distortion tradeoff is clear: more, smaller classes raise complexity but lower distortion (because each datapoint is closer to its class center). The general term "Bayesian" is often informally applied to most or all of these mathematically similar and philosophically related methodologies (McLachlan & Basford 1988; Baxter & Oliver 1994).

Philosophical foundations aside, the major practical differences distinguishing this class of statistical methods from most other numerical clustering methods include:

- Classes may overlap in a probabilistic sense; a single point may typically be shared between two or more classes.

- The same objective function can be used to measure and guide both the particular class centers, variances, and memberships, and the total number of classes.

- The clustering process induces a generative model of the data, meaning the model can be used both to classify new points and to generate datasets of realistic-looking points (hence, in this case, plausible protein substructures).

Just as Bayesian and related statistical methods have gained prominence in sequence analysis, through Gibbs sampling (Lawrence & Reilly 1990) and Hidden Markov Models (Krogh *et al.* 1994), so have they come to prominence in structure motif discovery.

Hunter and States (1991) used AutoClass, a general-purpose Bayesian classification tool (Cheeseman *et al.* 1988), to discover structure motifs for amino acid segments of length five. Motifs are represented in this work by a probability distribution over Cartesian coordinates for the backbone atoms of each residue. They discovered 27 classes of structural motifs (in the highest likelihood classification), where the majority of coordinates were assigned with a high probability to a single class.

Another general purpose statistical classification and modeling tool, SNOB (Wallace & Dowe 1994), has also been used to discover novel local structure motifs (Dowe *et al.* 1996). In contrast to previous work, the researchers in this project recognized the superiority of circular von Mises distributions, as opposed to Gaussians, in modeling angular data such as backbone dihedrals. The SNOB program searches for MML-optimal classifications of objects which may be defined in terms of any number of real number, angular, or discrete attributes, each of which is modeled with the appropriate type of probability distribution.

The mutually-supervised sequence and structure networks designed by Lapedes, Steeg & Farber (1995) can be recast within a Bayesian unsupervised learning framework (Becker & Plumbley 1996). Indeed, other neural network learning methods, such as used in predicting aspects of protein structure (Kneller, Cohen, & Langridge 1990; McGregor, Flores, & Sternberg 1989), can also be understood within and often improved by a Bayesian analysis (MacKay 1992).

Conceptual Clustering

Conceptual clustering techniques share with their numerical counterparts the goal of partitioning the data into natural groupings. They have, however, an additional goal, which is to characterize the clusters in terms of simple and meaningful concepts, rather than in terms of a set of statistics. These methods predominantly use qualitative attributes. Some of these commonly considered in protein motif discovery are proximity and spatial configuration.

The term **concept formation** is normally used to refer to incremental conceptual clustering algorithms. Following the definition of Gennari, Langley and Fisher (1989), concept formation can be described as follows:

Given: a sequential presentation of objects and their associated description,

Find: 1) clusters that group these objects into classes; 2) a summary description (i.e., a concept) for each class; and 3) a hierarchical organization for these concepts.

Several useful concept formation algorithms currently exist, including UNIMEM (Lebowitz 1987) and Cobweb (Fisher 1987). These systems rely, however, on an object representation being expressed as a list of attribute-value pairs. This representation is not the most suitable for the domain of protein structure, where the most salient features of the object may involve relationships among its parts. An emerging area of interest in machine learning is the design of structured concept formation algorithms in which structure objects are formed and then organized in a knowledge base.

IMEM is a concept formation method specifically designed for objects or scenes described in terms of their parts and the interrelationships among these parts (Conklin & Glasgow 1992; Conklin *et al.* 1996). These relationships may be topological (e.g. connectivity, proximity, nestedness) or spatial (e.g. direction, relative location, symmetry). A molecular structure is represented in IMEM as an image, which comprises a set of parts with their 3D coordinates, and a set of relations that are preserved for the image. The IMEM algorithm uses an incremental, divisive approach to build a subsumption hierarchy that summarizes and classifies a dataset. This algorithm relies on a measure of similarity among molecular images that is defined in terms of their largest common subimages.

The IMEM approach has been implemented as a system to perform conceptual clustering with protein structure data (Conklin *et al.* 1996). Figure 3 illustrates a classification exercise for a given protein fragment. Assuming the initial hierarchy of Figure 3a, the fragment in 3b is initially stored as a child of the most specific subsuming motif. The result of this step is pictured in Figure 3c. A concept formation step then occurs where a novel motif, which subsumes both the new fragment and a previously classified fragment, is generated. This last step is illustrated in 3d.

The Cobweb system (Fisher 1987) performs incremental, unsupervised concept formation using an information-theoretic evaluation function to construct a concept hierarchy. Specialized versions of Cobweb and AutoClass (Cheeseman *et al.* 1988) have been used to classify pairs of secondary structure motifs in terms of super-secondary motifs (Schulze-Kremer & King 1992). The results of applying these clustering methods were combined with results to form a consensus clustering.

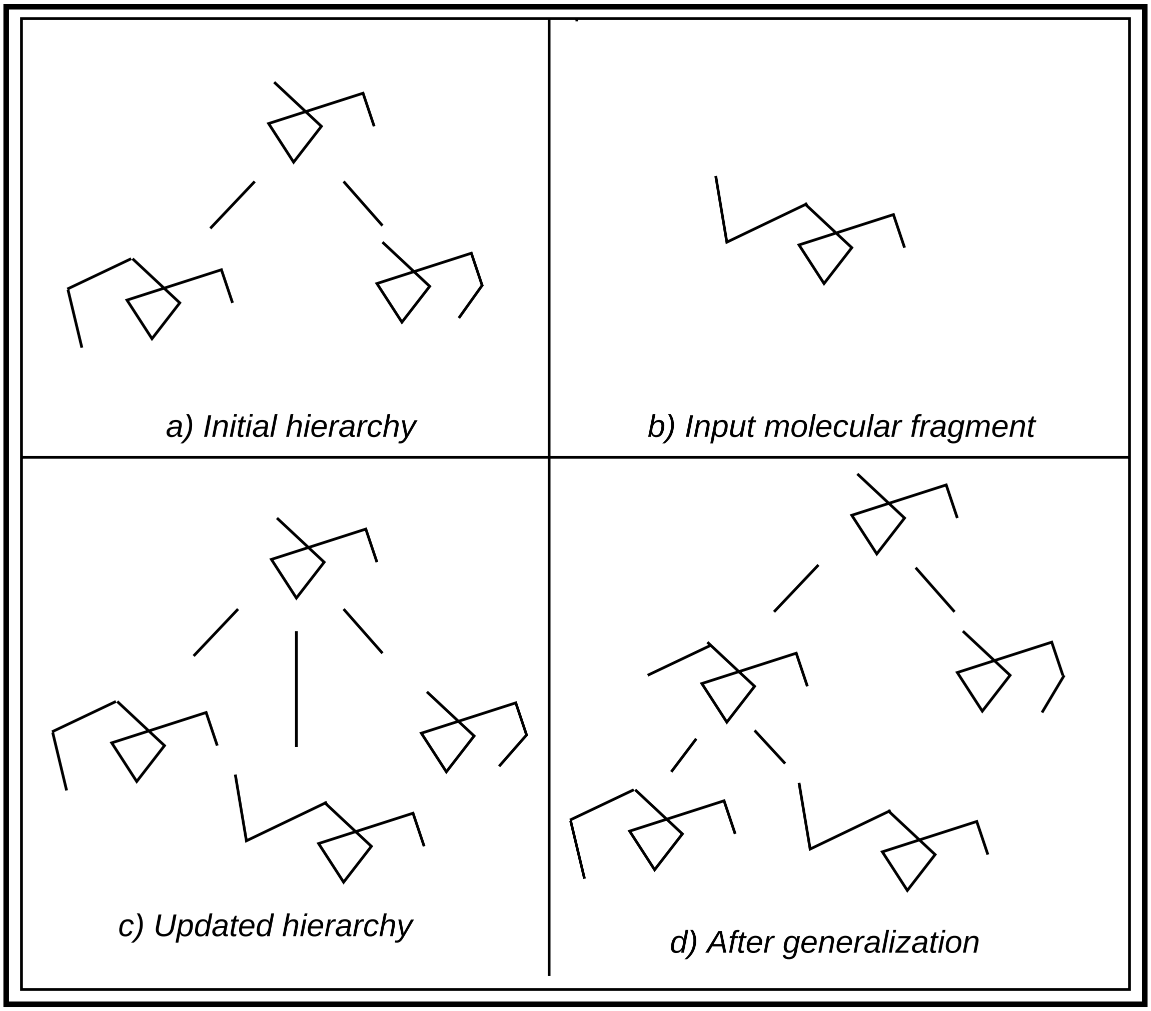

Figure 3: Development of a concept hierarchy for molecular images

Conceptual clustering has also been applied to sequence motif discovery. Shimozono *et al.* (1992) developed a decision tree method to classify membrane protein sequences according to functional classes using a variation of the conceptual learning system ID3 (Quinlan 1986).

Non-Clustering Techniques

There are a number of techniques that are not based on clustering *per se* but which also discover, or support the discovery of, protein structure and structure-sequence motifs. Prominent among these are 1) methods based on graph-matching that find recurrent motifs in terms of maximal common substructure or similar combinatorial criteria; 2) feature-selection methods that may be viewed as performing an essential first step in motif discovery; and 3) methods for finding empirical folding potentials.

Koch, Lengauer and Wanke (1996) used a variation of maximum clique enumeration to find maximal common subtopologies in a set of protein structures. They represented protein secondary structures as unlabeled directed graphs and restricted the usual clique enumeration algorithm to only those cliques that represent connected substructures.

It is important to ensure that the right information is input into a motif discovery process if one expects meaningful output. Feature discovery and feature selection are often crucial initial steps within a larger motif discovery task. In the protein structure analysis domain, many primitive features are selected by hand, based on prior expert knowledge. Other features may be discovered as special subsets or combinations of hand-picked features. Principal component analysis is commonly employed for this task on vector representations, because it can take an n-dimensional representation of a dataset and return a reduced representation in terms of $m < n$ decorrelated dimensions. This has advantages in later stages of machine learning or other processing. Other, analogous methods may be used on non-vector representations. For example, algorithms exist for finding pairs (Klingler & Brutlag 1994; Korber *et al.* 1993) or k-tuples (Steeg 1997) of correlated residue positions in sets of aligned protein sequences. Such correlations may bespeak evolutionarily-conserved structural or functional interactions, and form the basis for a particularly useful kind of structure-sequence motif. The discovery of correlations and associations in protein representations need not be limited to residues alone; one group of researchers has used a general-purpose data mining tool (Agrawal *et al.* 1996) to find substructures that correlate with particular functional features (Satou *et al.* 1997).

Another important type of protein motif is the empirical folding potential. Sippl has put this expanding enterprise on firm theoretical ground by using an Inverse Boltzmann's Equation to translate between theoretical force field terms and empirical database frequencies (Sippl 1990). By performing statistical studies of amino-amino proximity relationships, core/surface preferences of particular amino acids, and so on, Sippl and others (Grossman, Farber, & Lapedes 1995; Jones, Taylor, & Thornton 1992) have built motifs into objective functions that can supplement or replace theoretical potentials in structure prediction and threading.

Assessing Protein Motifs and Discovery Methods

A reading of the relevant research literature in protein analysis suggests at least six broad criteria by which to measure protein motifs, and, by extension, the methods used to define and discover them:

1. Predictability is the degree to which a motif representing one level or facet of protein structure or function may be predicted from knowledge of another. For the local structure motifs designated as "secondary structure", predictability is the ability to accurately predict secondary structure classes from amino acid sequence.

2. Predictive Utility is the flip side of the predictability criterion. For example, if one takes the view of secondary structure as an intermediate-level encoding between primary structure (sequence) and tertiary structure, then predictive utility ought to be some measure of the gain in accuracy in predicting tertiary structure with a particular encoding, as compared with prediction using other possible encodings. Another more direct measure might be the degree to which a particular set of proposed motifs, corresponding to secondary structure classes, constrain the ϕ and ψ angles of the included structure fragments.

3. Intelligibility refers to the ease with which researchers and practitioners of protein science can understand a given structure motif and can incorporate its information into their own work. Many factors affect intelligibility. For example, a discovered structure class that contains one-third traditional α-helix, one-third traditional β-sheet and one-third coil is harder to explain than one which correlates almost perfectly with α-helix. Also, for example, a motif expressed in first-order logic with terms for well-known biochemical aspects such as amino acid names and dihedral bond angles is easier to understand than a motif represented only in a set of several hundred neural network connection weight values. Further aspects of motif intelligibility are discussed below.

4. Naturalness, or the equally unwieldy word "intrinsicness", means the degree to which a motif captures some essential biochemical or evolutionary properties, or some essential class structure in the space of protein sequence or structure fragments under consideration. Some clustering methods are infamous for finding ersatz clusters in uniformly distributed data. Other clustering methods produce results very dependent upon their starting point. To avoid such results it is important to carefully choose appropriate representations and attributes for classification.

5. Ease of discovery refers to the computational complexity and data complexity of the methods required to discover the motif.

6. Systematicity is the degree to which a motif discovery method is derived from explicitly-stated principles and the degree to which the method can repeatably be applied to diverse data and produce consistent results.

Issues in Protein Motif Discovery

What kinds of issues determine the desirability of particular motifs and methods under the criteria listed above? Following, we present a discussion of some of the important considerations for carrying out a motif discovery exercise.

Use of Both Sequence and Structure Data

The discovery of sequence motifs has been, along with its associated goal of multiple sequence alignment, the mainstay of computational molecular biology from the beginning. A central hypothesis is that similarity of sequence implies similarity of structure and function, and evolutionary conservation of sequence implies conservation of structure and function. Thus, it is in a sense difficult to find any work on sequence motifs that does not involve sequence-structure motifs. Much of the vast literature on consensus sequences, sequence profiles, and Hidden Markov Modeling of protein families describes the discovery of motifs over sets of sequences that are already known to correspond to one particular overall tertiary structure. But where researchers have sought to define family-independent subsequence motifs that carry important structural information, they have not necessarily succeeded. In a study described in (Rooman, Rodriguez, & Wodak 1990), 11 out of a set of 12 sequence-structure motifs claimed to be predictive of secondary structure were found not to be.

There are two major limitations to the above methods. First, the sequence information is incorporated into the structure motifs after the latter have been defined. This approach is not designed to, and is not likely to, produce structure classes predictable from amino acid sequence. Second, as noted by Conklin (Conklin 1995b), in the methods outlined above, each structure motif may be associated with only a single sequence motif; there is no provision made for associating a structure motif with a disjunction of different sequence motifs (except in the narrow disjunction implicit in the abstraction of several very similar sequences into a more abstract motif).

The IMEM method proposed by Conklin, Fortier & Glasgow (1993) addresses both of these limitations by representing both sequence and structure objects in the symbolic format of a spatial description logic, a restricted first-order logic used to describe and manipulate concepts. Motif discovery in this system occurs through similarity-based clustering (structured concept formation) of combined sequence-structure representations. The sequence-structure predictability is built in to the discovery process, and enforced through a series of tests which measure the strength and significance of associations between sequence and structure motifs. For example, the ratio M_+/N – where N is the number of protein fragments assigned to the sequence portion of the motif, and M_+ is the number of fragments assigned to the joint structure-sequence motif – must be greater than 0.8. This ensures that more than 80% of all the instances of the sequence motif have the same structure, and therefore that the sequence may be predictive of the corresponding structure. Another test is a χ^2 test applied to a 2-by-2 contingency table for each structure-sequence motif. This test assesses the significance of association, over all protein fragments in the dataset, between the sets of fragments assigned or not assigned to the sequence portion of the motif and the sets of fragments assigned or not assigned to the structure portion.

Number of Classes, Number of Motifs

The number of different motifs, or classes, sought in a discovery procedure has important impact on both the information-theoretic aspects of predictive utility of the resulting motifs and the general intelligibility and usefulness of the motifs to molecular biologists.

First, in terms of the distortion versus complexity tradeoff in clustering and latent class modeling it is clear that more classes generally imply lower distortion and higher complexity costs. That is, the larger the number of classes and hence class centroids (exemplars, control points), the closer a given point will be to the centroid of its own class, ceteris paribus. But the larger the number of classes is, the more bits it takes to encode each data point in terms of its class-label encoding.

Though the MDL communications paradigm (minimizing a total number of bits transmitted between an hypothesized Sender and Receiver) is a somewhat artificial theoretical tool, it does reveal important practical aspects of data models. As the number of classes in a model of structure fragment data increases, a real tradeoff becomes apparent. Each motif becomes more specific, in that it carries more detailed local structural information about a smaller set of fragments. This might make subsequent tertiary structure prediction easier, because structure-packing considerations are made more precise. On the other hand, there is a loss of abstraction, a greater number of parameters to optimize in the motif discovery algorithm, a potentially greater difficulty in finding statistically significant estimates of frequencies and probabilities of motifs and features.

A growing consensus in computational molecular biology favors classes less coarse than the standard $2-5$ secondary structure classes. Conklin, in his survey (Conklin 1995b) cites three reasons:

1. First, he claims that there exist wide discrepancies between different methods of assigning secondary structure designations from crystallographically determined structures. This point is debatable. It appears to other observers that the Kabsch and Sander standard (1993) is both well-founded and widely accepted. However, to the extent that discrepancies do exist, one must take care that a prediction system is not just modeling the idiosyncrasies of particular structure definition rules.

2. A great number of fragment patterns tossed into the large default class "random coil" are neither random nor undefinable. Add to this the fact that different kinds of helices, and different kinds of β-strand configurations, can be observed, and there is a case to be made for additional subclasses of the three major classes.

3. Secondary structure packing analysis is a non-trivial task, and more accurate descriptions of local backbone structure – as ought to result from motif discovery with larger numbers of classes – can make the task much easier.

Locality: Size of Input Fragments

The size of sequence and structure fragments input to motif discovery systems is another issue closely related to the question of abstraction versus specificity. Smaller fragments imply smaller, more localized motifs. Smaller motifs mean that a greater number of them are needed to represent an entire sequence or structure, and hence a greater number of parameters are used in latter stages of a modeling or prediction task. On the other hand, smaller motifs also correspond to more frequently-occurring patterns, and therefore problems in probability estimation are minimized.

One must also carefully consider domain-specific and goal-specific criteria when choosing fragment size: Over what lengths of sequence and of structural backbone chain do the phenomena of interest manifest themselves? For example, individual β-strands can be captured with fragments of size 6 to 12, typically; but what about the turns between strands? How much information about the strand is conveyed by the nearby turns, and vice versa? How much information do different strands carry about each other? How much non-local information is necessary to determine a sequence fragment's propensity to "become" a helix or a strand or a stretch of coil, for example?

The information-theoretic and the biophysical issues here are deep and complex. An empirical, trial-and-error approach might be reasonable in attacking this problem. For example, a fragment size (prediction "window" size) of 13 was found to be effective in previous work on secondary structure prediction (Qian & Sejnowski 1988; Kneller, Cohen, & Langridge 1990). In such studies it was found that smaller windows failed to provide sufficient local contextual information for prediction of the secondary structure of the central residue in the window; for windows of length larger than 13, the marginal gains in extra contextual information were swamped by noise.

Another issue is fixed- versus variable-length motifs. For reasons described earlier in this chapter, most of the reported projects in structure-sequence motif discovery looked for short, fixed-length motifs of size 5 to 8 (Zhang & Waltz 1993; Hunter & States 1991; Unger *et al.* 1989). It has been observed, however, that different phenomena manifest themselves over different lengths (such as helices and turns), and different pieces of a protein have evolved and were conserved over different lengths of 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. Thus, it is too restrictive to discover motifs of only one size. An advantage of the IMEM approach (Conklin *et al.* 1996) is that is can discover variable-length motifs.

Representation

Perhaps the most important initial choice to be made in designing a motif-discovery method is what kind of representation to use for motifs. The differences between some of the options are huge, as large as the traditional gulf between the "symbolic/logical" and "numeric/statistical" camps in artificial intelligence research. The stakes can also be high, both in terms of the amount of interesting information captured by the resulting motifs and in terms of the ability for us to understand and communicate the information.

For structure classification, numerical clustering methods dominate the field (Zhang & Waltz 1993; Hunter & States 1991; Rooman, Rodriguez, & Wodak 1990). There are good reasons for this. First, structures are geometric and physical objects, and the representation of such objects – in terms of vectors, angles, and chemical properties – is an old and strong tradition in the physical and computational sciences. Second, the use of numeric features and statistical clustering techniques is very amenable to the use of well-defined objective functions, thus enabling a generally principled approach and the use of well-understood optimization procedures.

If the goal is intelligibility of derived motifs, there is no contest — logical representations are preferred. Clearly it is difficult to look at a set of hundreds of connection weights or means and variances and see anything resembling a motif. However, once a set of classes has been discovered, there is no major obstacle to finding more recognizable and descriptive motifs after the fact. The set of sequences, for example, corresponding to a particular structural class can be aligned, clustered, and so on, using standard methods, and consensus sequences can be produced.

A virtue of the IMEM method for representations (Conklin, Fortier, & Glasgow 1993) is that sequence and structure motifs are represented using a common formal syntax. Unlike the other structure-sequence motifs mentioned in this chapter and surveyed in (Conklin 1995b), the structure-sequence motifs of IMEM do not inherit a "confused dual semantics", where a sequence is interpreted using one formalism and a structure another. The knowledge representation formalism implemented in IMEM presumably enables it to be integrated more easily into larger, multi-level, multi-view protein analysis systems wherein many different kinds of features are used to predict other features.

Although a logic representation may be preferable, it is often difficult to determine the appropriate primitive concepts and qualitative relationships necessary for conceptual clustering. An integrated approach that incorporates both numeric and conceptual methods could address this issue. As a first step, numeric techniques could be used to perform an initial classification and derive parameters to be applied in a second step that would use conceptual clustering to derive meaningful (logical) concepts for the discovered motifs.

Intrinsic versus Extrinsic Clustering Criteria

Implicit in some of the above discussion is a concept of intrinsic versus extrinsic criteria for clusters and motif discovery. In the multi-stage process and multiple levels of description that characterize protein structure prediction, as in machine vision and speech recognition, there is a tension between the "best" intermediate representation language suggested by optimizing local, "current-level" criteria (What are the best clusters in $\Phi\Psi$ space?) and those suggested by optimizing "next-level-up" criteria (Which clusterings produce classes that work well as primitive symbols in a tertiary structure encoding?) This is a fundamental issue not yet addressed sufficiently in the computational molecular biology domain. There may be insights to be gained by examination of other domains which require motif discovery at several levels of organization, such as computer vision, speech recognition and natural language understanding.

Summary

The investigation of relations between protein tertiary structure and amino acid sequence is of enormous importance in molecular biology. The automated discovery of recurrent patterns of structure and sequence is an essential component of this investigation. This chapter has provided an overview of existing methods for protein structure motif discovery and some of the outstanding issues involved in this field.

Traditional machine learning and knowledge discovery techniques are not always appropriate for learning in the protein structure domain. This is mainly because they assume that similarity between objects is measured as a distance function based on simple attributes and values. The representation issues for structure motifs, however, are more complex; similarity is often judged in terms of spatial relations among parts of a structure as well as in terms of attributes of the structure and its parts. Another distinction is the size of existing datasets and the implied efficiency considerations resulting from the vastness and complexity of the data. Thus, there is an ongoing challenge to find appropriate methods for gathering information and knowledge from the evergrowing repositories of protein data, and in particular for understanding the intricate relationship between sequence and structure data.

Defining Terms

agglomerative. A clustering technique that incorporates a starting point that consists of as many clusters as there are instances in the dataset.
amino acid residue. The monomeric units of a protein. All proteins are composed of 20 standard amino acids that are linked together by peptide bonds.
clustering. The organization of data so that related information is together. There are several approaches to clustering, including numerical, statistical and conceptual.
concept formation. A term that refers to incremental conceptual clustering algorithms.
divisive. A clustering technique that incorporates a starting point that consists of a single cluster.
Hidden Markov Models. A statistical method developed for the generation, recognition and alignment of sequential data.
incremental clustering. A clustering technique for which instances are considered sequentially.
knowledge discovery. A field of artificial intelligence that involves computational theories and tools for assisting humans in extracting useful information from databases.
inverse protein folding. The process of fitting a known structure to a given sequence (rather than the more standard practice of trying to predict the structure of a protein from the sequence).
motif. An abstraction over a set of recurring patterns.
nonincremental clustering. A technique where all instance are considered at the outset of clustering.
secondary structure. The local conformation of the protein backbone. The most common folding patterns are helices, sheets and turns.
sequence motif. A linear string of amino acid residues (or residue classes).

sequence-structure motif. A sequence motif that associates each residue in its string with a particular secondary structure.
structure motif. A 3D graph where nodes on the graph correspond to a portion of the protein backbone.
a **structure-sequence motif.** A structure motif for which nodes in the graph are annotated with residue names or properties.
super-secondary structure Recurring groupings of secondary structure units (e.g., $\alpha\alpha$ unit, $\beta\alpha\beta$ unit, β barrell).
tertiary structure. The global conformation of a protein folded in 3D space.

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast discovery of association rules. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* American Association for Artificial Intelligence. 307–328.

Bairoch, A. 1991. Prosite: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* (19):2241–2245.

Baxter, R. A., and Oliver, J. J. 1994. MDL and MML: Similarities and differences (introduction o minimum encoding inference – part iii). Technical Report 207, Department of Computer Science, Monash University, Clayton, Vic. 3168, Australia.

Becker, S., and Plumbley, M. 1996. Unsupervised neural network learning procedures for feature extraction and classification. *International Journal of Applied Intelligence* 6(3).

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

Brachman, R. J., and Anand, T. 1996. The process of knowledge discovery in databases. In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* AAAI Press/The MIT Press. 37–57.

Buhmann, J., and Kuhnel, H. 1993. Complexity optimized data clustering by competitive neural networks. *Neural Computation* 5:75–88.

Cheeseman, P.; Kelly, J.; Self, M.; Stutz, J.; Taylor, W.; and Freeman, D. 1988. Autoclass: A Bayesian classification system. In *Fifth International Conference on Machine Learning.*

Chou, P. Y., and Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology* 47:45–147.

Conklin, D., and Glasgow, J. 1992. Spatial analogy and subsumption. In Sleeman, and Edwards., eds., *Machine Learning: Proceedings of the Ninth International Conference ML(92)*, 111–116. Morgan Kaufmann.

Conklin, D.; Fortier, S.; Glasgow, J.; and Allen, F. 1996. Conformational analysis from crystallographic data using conceptual clustering. *Acta Crystallographica* B52:535–549.

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.* AAAI/MIT Press, Menlo Park, California.

Conklin, D.; Fortier, S.; and Glasgow, J. 1994. Knowledge discovery of multilevel protein motifs. In Altman, R.; Brutlag, D.; Karp, P.; Lathrop, R.; and Searls, D., eds., *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press.

Conklin, D. 1995a. *Knowledge Discovery in Molecular Structure Databases.* Ph.D. Dissertation, Department of Computing and Information Science, Queen's University.

Conklin, D. 1995b. Machine discovery of protein motifs. *Machine Learning* 21.

Dowe, D.; Allison, L.; Dix, T.; Hunter, L.; Wallace, C. S.; and Edgoose, T. 1996. Circular clustering of protein dihedral angles by minimum message length. In *Pacific Symposium on Biocomputing '96*, 242–255.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds. 1996. *Advances in Knowledge Discovery and Data Mining.* American Association for Artificial Intelligence.

Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery: An overview. In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining.* AAAI Press/The MIT Press. 1–34.

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

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.

Gennari, J.; Langley, P.; and Fisher, D. 1989. Models of incremental concept formation. *Artificial Intelligence* 40:11–61.

Glasgow, J.; Fortier, S.; and Allen, F. 1993. Molecular scene analysis: crystal structure determination through imagery. In Hunter, L., ed., *Artificial Intelligence and Molecular Biology.* AAAI Press, Menlo Park, California. 433–458.

Grossman, T.; Farber, L.; and Lapedes, A. 1995. Neural net representations of empirical protein potentials. In Rawlings, C.; Clark, D.; Altman, R.; Hunter, L.; Lengauer, T.; and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology.* AAAI Press.

Hunter, L., and States, D. 1991. Applying Bayesian classification to protein structure. In *Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications.*

Jarvis, R. A., and Patrick, E. A. 1973. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers* C-22(11):1025–1034.

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

Jones, C. A., and Thornton, D. 1994. Protein superfamilies and domain superfolds. *Nature* 631–634.

Jones, D.; Taylor, W.; and Thornton, J. 1992. A new approach to protein fold recognition. *Nature* 358 86-89.

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

Klingler, T. M., and Brutlag, D. L. 1994. Discovering structural correlations in alpha-helices. *Prot. Sci.* 3:1847–1857.

Kneller, D. G.; Cohen, F. E.; and Langridge, R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214:171–182.

Koch, I.; Lengauer, T.; and Wanke, E. 1996. An algorithm for finding maximal common subtopologies in a set of protein structures. *J. Comput. Biol.* 2:289–306.

Korber, B.; Farber, R.; Wolpert, D.; and Lapedes, A. 1993. Covariation of mutations in the V3 loop of HIV-1: An information-theoretic analysis. *Proc. Nat. Acad. Sci.* 90.

Krogh, A.; Brown, M.; Mian, I. S.; Sjolander, K.; and Haussler, D. 1994. Hidden Markov Models in computational biology: Applications to protein modeling. *J. Mol. Biol.* 235:1501–1531.

Lapedes, A.; Steeg, E.; and Farber, R. 1995. Use of adaptive networks to evolve highly predictable protein secondary-structure classes. *Machine Learning* 21:103–124.

Lathrop, R. H., and Smith, T. F. 1995. Global optimum protein threading with gapped alignment and empirical pairing score functions. *J. Mol. Biol.* 255.

Lathrop, R.; Webster, T.; Smith, R.; Winston, P.; and Smith, T. 1993. Integrating AI with sequence analysis. In Hunter, L., ed., *Artificial Intelligence and Molecular Biology.* AAAI Press/ The MIT Press. 211–258.

Lawrence, C., and Reilly, A. 1990. An expectation maximization (EM) algorithm for the identificationand char acterization of common sites in unaligned biopolymersequences. *Proteins* (7):41–51.

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

MacKay, D. J. C. 1992. *Bayesian Methods for Adaptive Models.* Ph.D. Dissertation, California Institute of Technology.

McGregor, M.; Flores, T.; and Sternberg, M. 1989. Prediction of beta-turns in proteins using neural networks. *Protein Eng.* (2):521–526.

McLachlan, G., and Basford, K. 1988. *Mixture Models. Inference and Applications to Clustering.* Marcel Dekker, Inc.

Orengo, C. A.; Jones, D. T.; and Thornton, J. 1994. Protein superfamilies and domain superfolds. *Nature* 372:631–634.

Pauling, L.; Corey, R.; and Branson, H. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* (37):205–211.

Ponder, J., and Richards, F. 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.

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

Quinlan, J. 1986. Induction of decision trees. *Machine Learning* 1:81–106.

Rooman, M., and Wodak, S. 1988. Identification of predictive sequence motifs limited by protein structure data base size. *Nature* 335:45 – 49.

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

Satou, K.; Shibayama, G.; Ono, T.; Yamamura, Y.; Furuichi, E.; Kuhara, S.; and Takagi, T. 1997. Finding association rules on heterogeneous genome data. In *Pacific Symposium on Biocomputing '97*, 397–408.

Schulze-Kremer, S., and King, R. 1992. IPSA - inductive protein structure analysis. *Protein Engineering* 5(5):377–390.

Shimozono, S.; Shinohara, A.; Shinohara, T.; Miyano, S.; Kuhara, S.; and Arikawa, S. 1992. Finding alphabet indexing for decision trees over regular patterns: An approach to bioinformatical knowledge acquisition. Technical report, Research Institute of Fundamental Information Science Report RIFIS-TR-CS-60, Kyusha University.

Sippl, M. J. 1990. The calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures of globular proteins. *J. Mol. Biol.* 213:859–883.

Steeg, E. W. 1997. *Automated Motif Discovery in Protein Structure Prediction.* Ph.D. Dissertation, Department of Computer Science, University of Toronto.

Taylor, W., and Thornton, J. 1984. Recognition of super-secondary structure in proteins. *J. Mol. Biol.* (173):487–514.

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

Thornton, J., and Gardner, S. 1989. Protein motifs and data-base searching. *Trends Biochem. Sci.* (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.

Wallace, C., and Dowe, D. 1994. Intrinsic classification by MML - the SNOB program. In *Proc. 7th Australian Joint Conference on Artificial Intelligence*, 37–44.

Willett, P. 1987. *Similarity and Clustering in Chemical Information Systems.* Letchworth, Herts., U.K.: Research Studies Press.

Wilmot, C., and Thornton, J. 1988. Analysis and prediction of the different types of beta-turns in proteins. *J. Mol. Biol.* 203.

Zhang, X., and Waltz, D. 1993. Developing hierarchical representations for protein structures: An incremental approach. In *Artificial Intelligence and Molecular Biology.* AAAI Press (MIT Press).

Zhang, X.; Fetrow, J.; Rennie, W.; and Waltz, D. 1993. Automatic 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.* AAAI/MIT Press, Menlo Park CA.

Further Information

Recent research in the area of knowledge discovery is surveyed in the book *Advances in Knowledge Discovery and Data Mining* (Fayyad *et al.* 1996).

A good overview of the area of sequence motif discovery can be found in (Lathrop *et al.* 1993). The theses of Evan Steeg (Steeg 1997) and Darrell Conklin (Conklin 1995a) include more complete reviews of statistical and conceptual clustering techniques and their application to protein structure motif discovery.

The *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* (ISMB) published by AAAI/MIT Press and the *Proceedings of the Pacific Symposium on Biocomputing* (PSB) published by World Scientific contain articles related to protein motif discovery.

The *Proceedings of the National Conference on Artificial Intelligence* (AAAI) and the *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, both published by AAAI Press, contain general articles on machine learning and motif discovery.

A special issue of the journal *Machine Learning* (volume 21, 1995) focussed on learning and discovery techniques for molecular biology.