CHAPTER

12

Molecular Scene Analysis: Crystal Structure Determination Through Imagery

Janice I. Glasgow, Suzanne Fortier & Frank H. Allen

1 Introduction

This chapter describes the design of a prototype knowledge-based system for crystal and molecular structure determination from diffraction data. This system enhances current methods for the determination and interpretation of protein structures by incorporating direct methods probabilistic strategies, experience accumulated in the crystallographic databases, and knowledge representation and reasoning techniques for machine imagery. The process of determining the structure of a crystal is likened to an iterative scene analysis,

---

This paper is based on "Crystal and Molecular Scene Analysis," by Glasgow, Fortier and Allen which appeared in the Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications, Miami Beach, Florida, Feb. 1991

which draws both from the long-term memory of structural motifs and the application of chemical and crystallographic rules.

A crystal consists of a regular three-dimensional arrangement of identical building blocks, termed the unit cell; a crystal structure is defined by the disposition of atoms and molecules within this fundamental repeating unit. A given structure is determined by interpretation of an electron-density image of the unit-cell contents which can be generated from the amplitudes and phases of the diffraction data. Normally, however, only the diffraction amplitudes can be measured experimentally: the necessary phase information must be obtained by other means. This is the classic *phase problem* of the crystallographic method.

The structure determination of small molecules (up to 150 or so independent non-hydrogen atoms) has become a routine process in the last fifteen years. This is best observed in the rapid growth of the Cambridge Structural Database which has seen its number of entries increase from 14,000 to 90,000 in that period of time. *Direct methods* have contributed much to this progress by providing a mathematical, computer oriented solution to the phase problem. By contrast, the determination of macromolecular structures remains a lengthy and difficult task in which the phase problem continues to be a major hurdle.

The initial electron-density images obtained for macromolecules are typically incomplete and noisy. Interpretation of these images often involves mental pattern recognition on the part of the crystallographer: the image is segmented into features which are then pattern matched against individual recollections of expected structural motifs. Once recognized, this partial structure information can be used to improve the phase estimates and hence the subsequent image. The success of this iterative approach to image reconstruction depends crucially on individual recall of existing structural knowledge and on the ability to recognize its presence in a noisy map.

Our proposed knowledge-based system incorporates databases of information on previously determined crystal structures from which templates for pattern matching can be derived. Clearly, this will enhance the memory capability of the individual crystallographer. Further, our approach to image reconstruction is influenced by some of the current cognitive theories for imagery. These theories suggest that an image is organized as a depiction of its meaningful parts and their spatial relationships. Based on this assumption, a schema has been designed and implemented in which an image is depicted as a multi-dimensional symbolic array. Here the symbols in the array correspond to the meaningful parts of the image. The schema also includes functions that correspond to the processes involved in mental imagery, functions which form the basis for effective pattern matching techniques.

In combination with the probabilistic direct methods, these concepts of knowledge-based imagery provide for a more fluid approach to crystal struc-

ture analysis: the phase determination is now guided by structural information established by a pattern recognition procedure which makes use of chemical and crystallographic reasoning. This allows the image reconstruction process to follow a hierarchical path and therefore take advantage of the structural organization of proteins.

In the next section we provide an overview of our symbolic array knowledge representation schema for imagery. In Section 3 we discuss how the crystallographic phase problem can be reduced to a search problem, and in Section 4 we describe the structural databases used in our system. Section 5 presents our iterative algorithm for crystal structure determination.

2 Imagery

Mental simulations can provide insights that contribute to effective problem solving techniques. James Watson reported visualizing "pairs of adenine residues whirling in front of my closed eyes" at a time when he and Crick were verging on solving the structure of DNA [Watson 1968]. Similarly, the chemist Kekulé reported that it was spontaneous imagery that led him to the discovery of the molecular structure of benzene [MacKenzie 1965].

In determining crystal structures, crystallographers also relate the use of mental visualization or imagery. The electron density representation of the unit cell contains features which must be interpreted in terms of the expected chemical constitution of the crystal. Additionally, the interpretation must conform to chemical and crystallographic rules, as established from earlier experiments. Thus, it is natural for crystallographers to use their own mental recall of known molecular structures, or of fragments thereof, to compare with and interpret the electron density features. Furthermore, since crystals are three-dimensional objects, this mental pattern recognition must involve the rotation and translation of images through space. Theories of cognition would support the view that humans do indeed perform these "mental rotation" functions [Shepard and Metzler 1971].

The schema we propose for crystal structure determination supports the functions of imagery and visualization by representing crystal structures using both descriptive and depictive knowledge representation techniques. A symbolic array data structure is used to denote the hierarchical and spatial structure of such an image. Information needed to construct a symbolic array as well as the chemical knowledge of a crystal are stored and manipulated as a frame structure.

In this section we describe a knowledge representation scheme for imagery that can be used to represent, manipulate and reason about crystal structures. Such a representation includes a data structure for storing an image and functions on the representation that correspond to the mental processes involved in imagery. Before presenting this scheme, we

overview some of the research in cognitive psychology that has contributed to its design.

2.1 Mental Imagery

After many years of neglect, the topic of mental imagery has recently emerged as an active area of research in cognitive psychology. A debatable issue in this research concerns the underlying representation of images [Block 1981]: is an image represented as a *description* or as a *depiction* of its components?

Those who support the descriptive approach in the imagery debate suggest that images are not a distinct domain and thus are represented and manipulated as propositions [Pylyshyn 1981]. Contrary to this, supporters of the depictive approach state that imagery does involve a unique class phenomena and that images are organized into meaningful parts that are represented in terms of their spatial relations. The descriptive approach to representing images is appealing since it provides an abstract representation without significant loss of information. Although this representation is sufficient, it has been argued that it may not always be the most desirable. By studying the way people make inferences concerning spatial relationships, [Kosslyn 1980] has argued that mental imagery is used extensively. Further, he argues that both descriptive and depictive representations of images are involved in these mental processes.

As an alternative to defining a formal model for imagery, Finke has summarized much of the research in mental imagery by defining a set of "unifying principles" [Finke 1989]. The *implicit encoding* principle states that imagery is used to retrieve information that was not explicitly stored in memory. The principle of *perceptual equivalence* suggests that imagery is functionally equivalent to perception, in the sense that similar mechanisms are activated when objects or events are imagined as when the same objects or events are perceived. The *spatial equivalence* principle states that an image preserves, though sometimes distorts, the spatial relations of objects. The principle of *transformational equivalence* proposes a similarity relation between imagined and physical transformations. The final principle, the *structural principle*, states that the structure of an image is coherent, well organized and can be reinterpreted. These five principles allow the underlying intuitions of imagery to be expressed and further developed without restricting a model to the point where it applies to only a single task.

The primary goal of research in machine imagery is to develop representational tools for building programs that reason about and solve difficult problems using imagery. Similar to the cognitive approach of Finke, in developing computational tools for imagery we do not wish to restrict ourselves to a single model. Rather, we define a representation that allows the broad principles of imagery to be captured and expanded on.

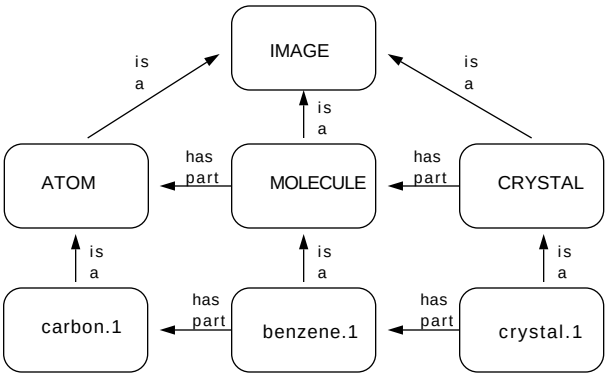

Figure 1. Semantic net for concepts and objects for crystallography domain

2.2 The Data Structure

The proposed representation schema for imagery is based on a formal theory of embedded, rectangular arrays [Jenkins and Glasgow 1989, More 1981]. Similar to set theory, array theory is concerned with the concepts of nesting, aggregation and membership. Array theory is also concerned with the concept of data objects having a spatial position relative to other objects in a collection. This theory strives to provide a universal recursive data structure that can be used effectively in a language that spans multiple programming paradigms. Such a language, Nial [Jenkins, Glasgow and McCrosky 1986], is being used to implement the computational processes for machine imagery as well as our knowledge-based system for crystal structure determination.

The array data structure provides a multi-dimensional realization of an image. The embedded nature of the array also allows for a parts-hierarchy depiction of an image. Detailed information (lower levels of the hierarchy) can either be hidden or made explicit in such a representation since hierarchical structure is expressed using embedded arrays. Thus the symbolic identification of meaningful parts of an image allows the depiction of a part to be suppressed unless attention is focused on that part.

Theories of cognition suggest that imagery involves both descriptive and depictive information [Kosslyn, 1980]. They have also suggested that in long-term memory depictive knowledge may be stored as a literal description of the locations of the parts within an image. Thus, the structure we propose is one which represents an image descriptively, yet allows a symbolic array depiction to be generated when needed [Glasgo and Papadias, 1992].

The important concepts in crystallography are: the periodically repeating motif in a unit cell of a crystal (CRYSTAL); the molecules and/or structural

FRAME: Dicyclohexano-18-crown-6 with potassium phenoxide and phenol	
class:	crystal
parts:	phenoxide-ring 0 0 0 DC-18-crown-6 1 1 1 phenol 2 2 2
molecular formula:	K-1 O-8 C-32 H-47
space group:	Pnca
unit cell dimensions:	14.15 23.794 9.491 90 90 90

Figure 2. Frame for a crystal structure

phenoxide ring		
	DC-18-crown-6	
		phenol

Figure 3. Two-dimensional projection of symbolic array for crystal structure

fragments (MOLECULE) and the atoms (ATOM). A semantic network that illustrates structural and hierarchical relationships between these concepts is presented in Figure 1. Instances of each of these concepts are also illustrated in the network.

A frame structure is used in our scheme to provide a descriptive representation of the concepts and objects for crystallography. An image frame has two required slots: a *parts* slot that provides a literal representation of the components of an image and their locations in Euclidean space; and a *depict* slot that contains a default function that generates the symbolic array representation of an image from the given parts and locations. An example of a frame that represents the depictive and descriptive knowledge of a crystal structure is illustrated in Figure 2.[1]

A symbolic array denotes the structural features of an image. This array may be depicted in one, two or three dimensions. Figure 3 illustrates a two-dimensional projection of the three-dimensional symbolic array data structure that would be generated using the depict slot of the image frame and the parts slot of the crystal frame.

Hierarchical organization is a fundamental aspect of imagery. Theories of selected attention suggest the need for an integrated spatial/hierarchical representation: when attention is focused on a particular feature, the brain is still

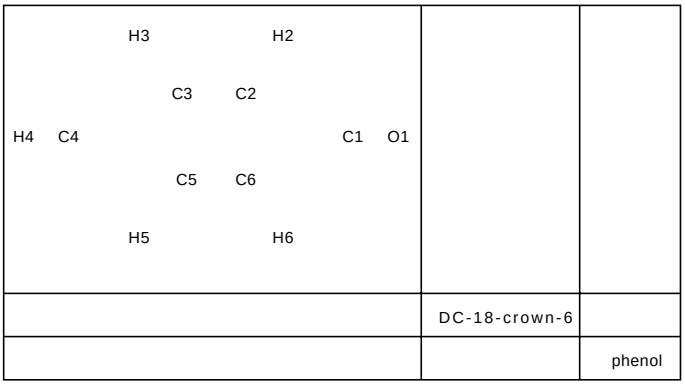

Figure 4. Embedded array representation of subimage phenoxide-ring, projected into two dimensions.

partially aware of other features and their spatial relation to the considered feature. Our symbolic array representation supports such theories by considering an image as a *recursive* data structure. A symbolic element of an array can itself denote a subimage. Consider the image of the crystal structure depicted in Figure 3. The symbols in this structured representation can denote structured subimages. Figure 4 illustrates the symbolic array depiction of the crystal structure when attention is focused on the subimage of phenoxide-ring. As with the image of the crystal, this embedded array would be generated using the depict slot of the image frame for the phenoxide-ring.

The primary goal of the knowledge-based system for crystal structure determination is to obtain a detailed and precise three-dimensional picture of the atomic arrangement in the crystal. In Section 5 we describe an algorithm that reconstructs such an image of a crystal in the form of a symbolic array.

A computational model for mental imagery has previously been proposed [Kosslyn 1980]. In his theory, images have two components: a surface representation (a quasi-pictorial representation that occurs in a visual buffer) and a descriptive representation for information stored in long-term memory. The two-dimensional surface representation of his theory is fundamentally different from the symbolic array representation described in our chapter. However, the design of the frame representation and the functions defined on images in our representation were greatly influenced by Kosslyn's empirical studies and model.

2.3 Functions on Images

The effectiveness of a scheme for knowledge representation is measured primarily by how well it facilitates the processes that operate on the representation. Larkin and Simon argue that diagrams are computationally prefer-

NAME	OPERATION
Retrieve	Retrieve an image representation from long-term memory.
Construct	Construct a symbolic array depiction of an image from a descriptive literal representation.
Compose	Compose two images into a single complex image with a given spatial relationship.
Symop	Use symmetry information to retrieve regularities in an image.
Resolve	Use pattern matching information and world knowledge to transform an image into one of higher resolution.
Compare	Compare images and determine similarity measure.
Consistent	Determine if an image is consistent with world knowledge.

Figure 5. Functions for constructing images

able to propositional representations, not because they contain more information but because the indexing of the information supports efficient computations [Larkin and Simon 1987]. In this subsection we propose a set of primitive functions that were designed to support the cognitive inferences involved in imagery. These functions on symbolic arrays are considered in three categories: functions for *constructing* images, functions for *transforming* images and functions for *accessing* images.

The first class of functions we consider are those involved in constructing an image. These functions are summarized in Figure 5.

Theories of cognition support three distinct memory systems: sensory storage, working or short-term memory and long-term memory [Baddeley 1986]. When considering machine imagery, we are mainly concerned with representation in working memory. One way to construct an image in working memory is to *retrieve* an instance of this image from long-term memory. Computationally, we interpret long-term memory as a database of frames that provide a propositional and a literal representation of an image. The creation of an image from such a representation is on an "if-needed" basis. An invocation of the function specified in the depict slot results in the *construction* of a symbolic array representation of the image from the literal representation of its parts.

A basic process of thought is the creation of new concepts from old. Imagery involves constructing and manipulating images in unique ways. Our model supports processes for imagery by allowing complex images to be constructed as a composition of simpler images. For example, we may store images of a ball and a box, but the image of a ball sitting on top of a box is created from the two subimages and the desired spatial relation. The hierarchical representation of images permits us to *compose* two or more images

NAME	OPERATION
Rotate	Rotate array depiction of an image a specified number of degrees around one of the axes.
Translate	Translate position of component within a symbolic array.
Zoom	Increase or decrease the apparent size of a depiction of an image.
Project	Project a three dimensional array onto two dimensions.

Figure 6. Functions for transforming images

into a single image in which the components are spatially related.

Images can also be constructed through the processes involved in perception and recognition. In this case the initial representation of an image comes from sensory store, which holds information impinging on the sense organs. As suggested in [Marr and Nishihara 1978], an image may go through several stages of representation going from perception to recognition. The processes involved in these transformations are complex and dependent on the domain. One such process is the ability to *compare* two images and determine a measurement of closeness. This measurement can depend on both spatial and non-spatial features of an image. We also consider the function *resolve* that takes the results of pattern matching and refines the image. Resolving an image may result from reconstructing recognized features into a new image. Once this has been done, the resulting image can be checked for consistency with world knowledge. In the crystallographic domain, for example, a complex image may be constructed that depicts molecules at too close a distance. Such an image could be evaluated as impossible by the *consistent* operator given this domain of interpretation.

Images are often constructed using incomplete knowledge. This is particularly true when considering three dimensional images where perception may be in two dimensions. In cognition, missing information can be provided by considering regularities such as symmetry in an image [Pentland 1986]. The function *symop* is used to retrieve regularities through symmetry operations such as reflection and rotation.

As suggested by empirical experiments in cognitive psychology [Shepard and Metzler 1971], mental imagery also involves processes for manipulating objects in space. Figure 6 summarizes the proposed functions for transforming images.

Note that the functions for transforming a symbolic array representation of an image are typically not meant as an alternative form of representation, but as a means of viewing an image from a variety of perspectives. These functions are necessary in developing a theory of recognition based on pattern matching of spatial images. For example, we may need to *rotate* and

NAME	OPERATION
Find	Scan a depiction of an image to determine the location of a component.
Focus	Shift attention to a particular component of an image.
Query	Retrieve propositional information from an array depiction of an image.

Figure 7. Functions for accessing images

translate a perceived image before it can be pattern matched with an image reconstructed from long-term memory.

The final class of functions corresponds to processes for accessing an image. These functions are particularly useful when reasoning about images. Figure 7 summarizes these functions.

The *focus* function is a mapping from an image and a symbolic feature of the image to a new image such that attention is concentrated on the specified feature. If the designated feature is not an atomic array, then the new image will be the old image with the feature replaced by its symbolic array representation. Knowledge involving the spatial relations of an image can be retrieved using the *query* operation.

2.4 Otherapplications of this work

The research in machine imagery underlying the crystallography system has impact beyond the molecular recognition application. Since the knowledge representation schema was designed to capture fundamental properties of mental imagery, its implementation provides a basis for the computational modeling of cognitive theories that involve processes related to imagery. These include the sequential and parallel processes involved in memory retrieval of images, attention, recognition, learning and classification [Glasgow 1990a, Glasgow 1990b]. As well, the scheme provides a framework for developing other knowledge-based applications. Currently, we are considering additional applications in the areas of haptic perception, game playing, medical imaging and robotic motion planning.

3 The Crystallographic Phase Problem

Even though the diffraction experiment yields several hundred to thousands of observations, and normally a relatively high ratio of observations to unknown parameters, the information sought - a three-dimensional picture of the atomic arrangement in the crystal - cannot be calculated directly from the measured data. This is because such a calculation requires knowledge of both

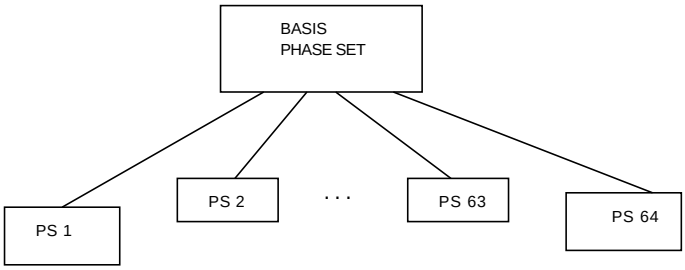

Figure 8. Phase Search Tree

the amplitude and the phase of each diffracted ray and the latter cannot be measured experimentally. This is a basic problem in any crystal structure determination, and it cannot be easily circumvented, despite the mathematical overdeterminacy, because of the Fourier transform relationship between the crystal structure and its diffraction pattern.

The most successful and straightforward approach to the solution of the phase problem are the so-called *direct methods* [Hauptman 1986]. This approach uses probability theory to retrieve phase information from the amplitude data. It essentially predicts the value of certain linear combinations of the phases and provides a way, through the variance of the distributions, of ranking the information according to reliability. The process evaluates several thousand of such linear equations and yields a redundant system of equations from which the values of the individual phases will then be extracted. Phasing is initiated from a basis set of, typically, four phases whose values can be selected (for origin and enantiomorph specification), together with a number of further selected phases whose values are permuted. This yields several possible solutions corresponding to the several possible phase permutation combinations, and indeed the method is referred to as the *multisolution approach.* Figures of merit are then calculated and used to assess which of the solutions appears to be the best one. The last, and finally the only important test, is whether or not any of the phase sets will produce an interpretable image of the structure.

The crystallographic phase problem can be thought of as a general search problem [Fortier, Glasgow and Allen 1991]. In this context, the multisolution approach can be described as a simple generate and-test search procedure as illustrated in Figure 8. The morphology of the search tree is unusual, though. The tree has a single depth level with a large branching factor; the number of nodes in such a tree is usually between 32 and 128. What characterizes and limits the search is the fact that the heuristic evaluation functions used (the

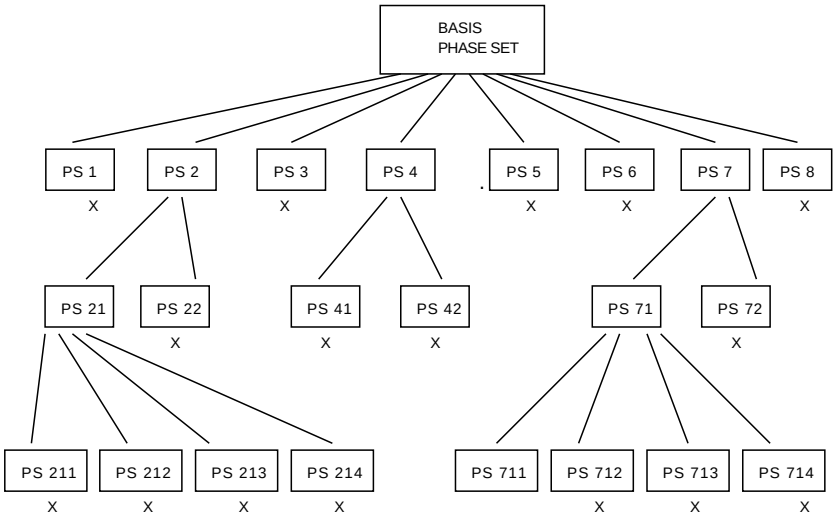

Figure 9. Hierarchical Phase Search Tree

figures of merit) are not effective in pruning partially developed solutions. This is because these heuristic functions do not test the actual goal of the search: the interpretability of the reconstructed image. Rather, they provide ranking numbers that are derived from the underlying probabilistic model. They depend, thus, on how well the actual structure fits the statistical model. The *a priori* model, assumed by direct methods, is that the repeating atomic motif in the crystal can be represented by a collection of atoms uniformly and randomly distributed through space.

Traditional direct methods explore the phase space and evaluate phasing paths by using only very general chemical constraints—the electron density distribution must be everywhere nonnegative and its peaks must correspond to atoms—and the constraints imposed by the amplitude data. While these constraints have proven sufficiently limiting for applications to small molecules, they are not adequate for tackling the more complex structures such as those of proteins.

Several additional chemical constraints, such as limits on bond lengths and angles or expected conformations, are usually available at the outset of a structure determination project. In a direct methods procedure, these additional constraints are supplied by crystallographers, who use their recall of existing results to reconstruct structural templates, and their visual abilities to match these templates with the developing electron density image. Recent theoretical results have shown, however, that information from partial struc-

ture identification, together with diffraction data, can be incorporated in a general direct methods joint probability distribution framework [Bricogne 1988, Fortier and Nigam 1989]. These results open the way for AI contributions to direct methods for solving protein structures by allowing for a flexible, context driven solution procedure which can automatically take advantage of all available information. In particular, the image reconstruction can be modeled as an iterative resolution process in which any structure recognition can be used to evaluate the phasing paths as well as guide the phasing exploration through the incorporation of further chemical constraints. A phase search tree that reflects this hierarchical strategy is illustrated in Figure 9. Furthermore, this work has provided the theoretical basis needed for the computer generation of the phasing distributions. It thus becomes possible to consider dynamic systems in which distributions, tailored to the knowledge base, are generated as needed.

4 The Structural Databases

Crystallographers, perhaps mindful of the intrinsic importance of their results, have a long and successful history of documentation. Early printed indexes and compendia have now been replaced by computer-based information banks. Complete three-dimensional structural data for some 140,000 compounds, from simple metals to proteins and viruses, are now stored in four crystallographic databases [Allen, Bergerhoff and Sievers 1987]. All of the databases are regularly updated with new material and the long-term memory of existing crystal structures increases by about 10% per year. The Cambridge Structural Database [Allen, Kennard and Taylor 1983] (CSD:90,000 + organo-carbon compounds) and the Protein Data Bank [Bernstein, Koetzle, Williams, Meyer, Brice, Rodgers, Kennard, Shimanouchi and Tasumi 1977] (PDB: 550 + macromolecules) contain the vast bulk of available experimental knowledge of three-dimensional molecular structures.

The systematic recall of three-dimensional structural knowledge from the databases is essential to our imagery approach to crystal and molecular structure determination. However, this knowledge is not explicit in the plethora of three-dimensional crystallographic facts, e.g. coordinates, cell dimensions, symmetry operators, etc., that dominate the information content of the databases. Rather, it must be derived from the stored facts via mechanisms for search, retrieval, classification, reasoning and rule generation. These activities are made possible by the rule-based two-dimensional representations of formal chemistry (structural diagrams or sequence data) that are also included in the databases. This is knowledge that can be searched using the syntactic language of chemistry and which underpins the interpretation of the three-dimensional structural facts.

A new crystal structure determination is seldom undertaken without some prior knowledge of the expected two-dimensional chemistry of the compound. This knowledge forms the basis for a database search to locate the key chemical fragments of the molecule, e.g. helices, rings, ring systems, acyclic functional groups, etc. In three-dimensions these fragments may occur in a number of different conformations, each of which represents a potential template for pattern matching with the electron density maps. Methods for machine learning, embodied in cluster analyzes based on suitable shape descriptors, serve to classify the database fragments into conformational subgroups. [Allen, Doyle and Taylor 1991]. The derivation of syntactic rules, which describe the conformational relationships of fragments one with another, provides a linguistic framework which permits larger templates to be built. Template generation and model building are active research areas for both small and large molecules (see e.g. [Dolata, Leach and Prout 1987; Wippke and Hahn 1988; Blundell, Sibanda, Sternberg and Thornton 1987]). The results, apart from their use in crystallography, are extensively used in rational drug design projects.

A systematic and comprehensive knowledge of the weak hydrogen-bonded and non-bonded interactions is also crucial to image reconstruction and validation from electron-density maps. These interactions not only govern the limiting contact distances between molecules in the crystal structure, but also play a key role in stabilizing the molecular structures of large molecules such as proteins and nucleic acids.

Crystallographic data have always been the primary source of information on the dimensions and directional preferences of hydrogen-bonded systems [Taylor and Kennard 1984]. The use of statistical analysis, decision theory, and the classification of H-bonded motifs observed in crystal structures, suggest rules that govern H-bond formation [Etter, MacDonald and Bernstein 1990]. Application of these rules provide knowledge of the environmentally dependent limiting geometries and motif templates that are relevant to crystal structure determination (see e.g., [Sawyer and James, 1982]).

The study of limiting contact distances between non-bonded atoms has a similar dependence on crystal structure results. Even today, most chemists use the non-bonded radii of Pauling [Pauling 1939] which are based on limited experimental data and assume that (a) non-bonded atoms are effectively spherical in shape and (b) the radii are additive and transferable from one chemical environment to another. Database analyzes (e.g. [Taylor and Kennard 1984, Allen, Bergerhoff and Sievers 1987]) indicate that these assumptions are inexact, i.e. the limiting contact distance between two atoms depends on their chemical environments and on their direction of mutual approach. Interest now focuses on geometries and motifs which are stabilized by even these weak intermolecular forces, both in small molecules [Desiraju 1989] and in proteins [Rowland, Allen, Carson and Bugg 1990]. The knowledge of

non-bonded atomic shapes defines not only the limiting contact distances between atoms, but also the spatial shape and size of both fragments and molecules. Further, the motifs formed by non-bonded interactions provide additional templates for use at all stages of a crystal structure determination.

In addition to statistical techniques, concepts from machine imagery research are being used for motif classification. Research in this area includes determining methods for developing classification schemes for images based on symbolic representations of both the conformation and configuration of molecules. Such a classification scheme will be used to extract syntactic rules of three-dimensional molecular structures from the crystallographic databases.

5 Crystal Structure Recognition

The problem of determining the structure of a crystal from diffraction data belongs to the general class of image reconstruction problems. The goal of the reconstruction is to produce a complete image which contains both depictive and descriptive knowledge of the three-dimensional atomic arrangement in the crystal. This image is built from information on the unit cell, the symmetry operators within the unit cell and finally the unique asymmetric portion of the repeating atomic motif. Determining a crystal structure is therefore analogous to a scene analysis in which the structural atomic motif enclosed in the unit cell is recognized by using a memory of previously determined structural templates and is understood through the application of chemical and crystallographic rules.

Thus our approach borrows partially from research in the area of vision. In particular, we incorporate existing segmentation algorithms that decompose an image into its meaningful parts. The technique used for recognizing a fragment of a crystal structure involves comparing its image to a stored representation of a previously recognized structure and evaluating the fit. This template matching approach is a simple and relatively old technique that has been used in vision applications. Recognition in our model also assumes constructed shape and volume descriptions, as in the approach of Marr [Marr 1982].

The crystallographic application differs from vision applications in a number of ways. First, the image for a crystal is perceived and depicted in three dimensions. This eliminates many of the problems of feature segmentation and recognition involved in vision applications: features in three dimensions do not overlap and we can utilize three-dimensional segmentation and pattern matching techniques. As well, we are not concerned with factors such as light sources, surface material or atmospheric conditions that may distort the appearance of a visual image. The complexity that does exist in the crystallographic application relates to the incompleteness of data due to

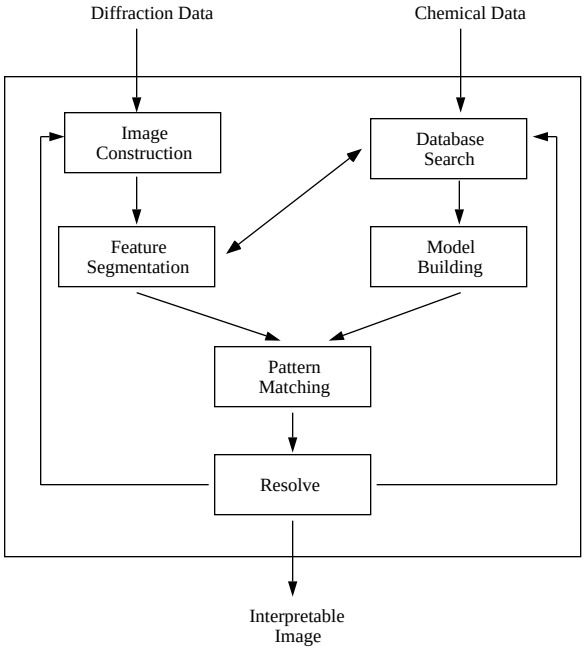

Figure 10. Algorithm for structure recognition

the phase problem.

In our approach, the process of crystal structure determination is modeled as *resolving* the three-dimensional image of the atomic arrangement within the crystal. By using a hierarchical approach, the phasing search space is expanded to a multilevel search tree. At each level of the search tree, any partial structure determined through pattern matching is used to update the probability distribution so as to provide higher resolution images. Thus, the identification process is an iterative one. Once an image of the crystal has been constructed, we focus on particular regions of the structure and try to pattern match them with structural templates from the database. Good matches serve not only to guide the identification search, but also to refine our image of these substructures and iteratively refine our complete structure.

Figure 10 illustrates the processes involved in structure recognition. Initially, an image of the structure, in the form of an electron density map, is constructed using the measured amplitudes and a given phase set. The known chemical information for the crystal is used to do a preliminary database search for fragments that could be anticipated in the structure. Simultaneously, the current image is segmented into distinct three-dimensional "blobs" or subimages that correspond to the structural features of the image. These fea-

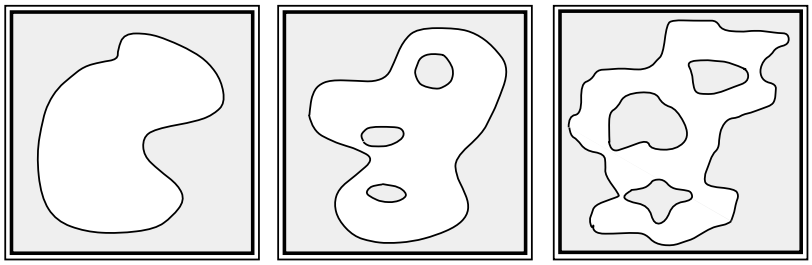

Figure 11. Resolution stages of molecular image

tures are then compared with images of the fragments retrieved from the database. Heuristics, based on the results of this three-dimensional pattern matching, are used to prune the search tree. In addition, matched images provide the necessary information for improving and expanding the phases and thereby resolving the image. These processes are repeated until the resolution of the image matches that of the diffraction data. Figure 11 illustrates a two-dimensional projection of images going through several stages of resolution, where the higher resolution images correspond to utilizing an increased phase set in their construction.

We now present a brief discussion of each of the steps in the crystallographic image reconstruction algorithm.

• **Image Construction.** Just as in vision, the image of a crystal may go through several stages of representation. At this step of the algorithm, the image of the crystal is represented as a three-dimensional electron density map resulting from the diffraction experiment and the current phase set. Figure 12(a) illustrates a two dimensional projection of a three-dimensional array representation of an electron density map, where the values in the array denote the electron density at the corresponding locations within the unit cell of a crystal.

Initially the electron density map is constructed using low resolution phases from the basis set expanded by selecting a small number of additional phases. Such a map will correspond to a low-resolution, noisy image of the crystal but, as additional phases are determined in successive iterations of the algorithm, the maps will reveal clearer and clearer (higher-resolution) images as illustrated in Figure 11.

• **Feature Segmentation**. In this process we partition the electron density map into distinct, three dimensional structural features. Standard image preprocessing techniques, such as noise reduction, local averaging, ensemble averaging, etc. are applied prior to segmentation. These techniques are used to enhance features of an image by establishing regions that either contain or

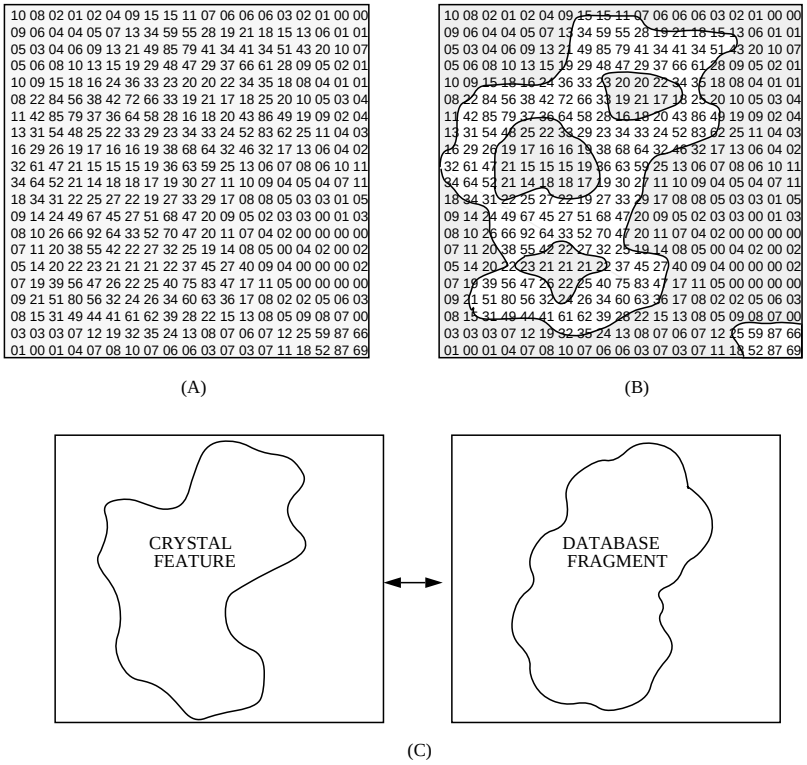

Figure 12. Two-dimensional projection of stages of image recognition: (a) Electron Density Map (b) Segmented Electron Density Map (c) Pattern matching step

do not contain electron density. A technique for determining distinct blobs/regions is then used to segment the map into features. World knowledge about anticipated shapes is used to determine whether these features are consistent with the chemical knowledge of the structure. Output from the process consists of a set of distinct blobs/regions that correspond to the structural features of the image; these may now be used to pattern match with anticipated patterns retrieved from the database. Figure 12(b) illustrates the blobs resulting from a segmentation process on an electron density map.

A library of segmentation functions for three-dimensional images is being implemented and tested on the images of crystal. Included in this library are functions that correspond to boundary detection, region growing, hierarchical and boundary melting techniques.[2]

We are also considering techniques that incorporate knowledge of the crystallographic domain. The selection of appropriate segmentation functions, to be used at each iteration of the algorithm, depends on the level of

resolution of the image being considered.

At this stage, we also determine some descriptive information about the segmented feature. This includes volume and shape information that can be used to assist in the pattern matching process.

• **Database Search.** The knowledge-based system is designed to incorporate information from the crystallographic databases described in Section 4. Prior knowledge of the two-dimensional chemistry of each new crystal structure defines a 'query' domain for a chemical search of the databases. This query is partitioned (a) to generate bonded chemical fragments for which likely three-dimensional templates are required for pattern matching, (b) to identify hydrogen-bond donors and acceptors present in the molecule, and (c) to identify atoms or functional groups which are likely to play a key role in the non-bonded interactions that stabilize the crystal and molecular structure. In (b) and (c) the databases are searched and analyzed to retrieve limiting geometries and likely three-dimensional motifs for use in pattern matching and image resolution.

• **Model Building.** Once an anticipated fragment has been retrieved from the database, a symbolic array image for the fragment is reconstructed. From this image of the fragment we can generate a blob-like depiction at a resolution level corresponding to the current resolution of the features of the crystal.

• **Pattern Matching.** The input to this process is the set of unidentified features derived from the segmentation step and the set of anticipated fragments selected from the database using chemical and structural information. The goal of the process is to compare each of the unidentified features with database fragments to determine the best three-dimensional structural matches. Both iterative and parallel algorithms for carrying out these comparisons are currently being considered [Lewis, 1990].

To facilitate a pairwise comparison, the three-dimensional representation of the known molecular structure is oriented within the cell of the unknown structure. Techniques from molecular pattern recognition are being used to achieve the correct position through rotation and translation [Rossman, 1990]. Patterson-based techniques are used to focus attention on the most promising regions of the electron density map. A template matching approach is then applied and the degree of fit assessed. Figure 12(c) illustrates a pair of subimages considered for pattern matching.

• **Resolve.** Information gathered from successful pattern matches (those with a high degree of fit) is used to update the phase set and subsequently generate a new electron density map for the crystal. This information is first checked for consistency with other knowledge for the domain; for example the image composition is checked against packing constraints for the crystal. The structural information, which is kept at a resolution level matching that of the current image, is then incorporated in the direct methods phasing tools.

This provides additional chemical constraints which serve in the improvement of the current phases and the expansion to higher resolution phases and therefore higher resolution images. Note that keeping the structure recognition information at the current image resolution level ensures that this information guides rather than drives the structure determination process.

The resolve process also controls the search space for the algorithm. Recall that we are attempting to reach a goal state in which enough phase information is available to construct an interpretable image. Incorrect pattern matches may lead to paths in the search tree (Figure 9) in which the expansion to higher resolution phases does not contribute to forming a clearer image. Intermediate evaluation functions applied to the evolving images allow us to prune such paths and only consider those that lead towards a goal state.

The processes described above are repeated until a fully interpretable image of the structure has been resolved. At this stage we can combine the *where* information, derived from the segmentation process, with the *what* information, derived from the pattern matching process, to construct a symbolic array representation for the crystal. The “where” information provides the exact location of each of the distinct features within the unit cell of the crystal; the “what” information gives the chemical identity of these features. By combining the “where” and “what” information in a symbolic array, we are able to reconstruct a precise and complete picture of the atomic arrangement for the crystal. Using the symbolic array representation and the known chemical data for the crystal, a frame representation can be constructed and added to the database of known structures.

Each individual module described above is being implemented and tested in an independent manner to establish an initial working prototype for each subtask. Once this preliminary, but extensive, work has been completed, the modules will be integrated and the system tested in its entirety. Currently, three-dimensional electron density maps are obtained by use of existing crystallographic software. A library of functions for the preprocessing and segmentation of these images at various levels or resolution is under development. As well, routines for the extraction of meaningful features (size, shape, centre of mass, etc.) from the derived segments are being developed. A prototype for the semantic network memory model has been established. The network incorporates the customory "chemical structure" hierarchy of protein structures (atom, residue, secondary structure,etc.) as well as the "classification" hierarchy that allows for the inheritance of properties. Routines for the construction of symbolic array and electron density map representations from the frame representations have been tested for selected cases. Work has begun on the pattern matching module and on the implementation of a direct space pattern matching function. The resolve module is still at the design stage, although several of the direct methods algorithms in its core have already been implemented and tested.

Although the initial implementation of the algorithm is sequential, the algorithm and the individual processes are being designed to incorporate any potential parallelism for later re-implementation. For example, we can concurrently process the pairwise pattern matching of fragments from the database with features from the crystal. Further, the hierarchical phase search tree (Figure 9) can be considered as an "OR" search tree. That is, if any path can be generated from an initial state to a goal state then a solution is found. Since these paths are independent, they can be generated in parallel.

In the reconstruction algorithm described above, imagery plays an important role in identifying crystal structures. The spatial/hierarchical structure of a crystal is represented as a symbolic array image. Such a representation can be transformed into three-dimensional depictions for pattern matching. Image transformation functions are then used to pattern match features of a crystal with the depictions of molecular structures reconstructed from the symbolic arrays. Ultimately, the image reconstruction process results in a symbolic array depiction for the initially unidentified crystal structure.

The programming language Nial, which is based on the theory of arrays, is being used to implement the prototype system. The array data structure and primitive functions of Nial allow for simple manipulations of the crystal lattice. Furthermore, the Nial Frame Language [Hache, 1986] provides an implementation for the frame structures used in the imagery model. Nial also provides the syntax to allow us to express the parallel computations inherent in our reconstruction algorithm [Glasgow, Jenkins, McCrosky and Meijer, 1989].

6 Related Work

Computer-assisted structure elucidation by use of knowledge-based reasoning techniques is one of the most active application areas of artificial intelligence in chemistry. When applied to two-dimensional structural chemistry, the goal is the interpretation of chemical spectra (mass spectra, IR, NMR data) in terms of candidate two-dimensional chemical structure(s). A number of systems have been developed, of which the DENDRAL project is by far the best known [Gray, 1986]. Some of the fundamental methodologies used in DENDRAL - for example, mechanisms and algorithms for knowledge representation, pattern matching, machine learning, rule generation and reasoning - have had a lasting impact on the computer handling of two-dimensional chemical structures. They have also contributed significantly to the development of chemical database systems and of tools for computer-assisted synthesis planning and reaction design (see e.g., [Hendrickson, 1990]).

Applications to three-dimensional structural chemistry and crystallography are still relatively new and comparatively more fragmentary. They can be broadly divided into two interrelated categories, depending on whether their main purpose is the classification or the prediction of three-dimensional

molecular structures. For small molecules, the primary application area is that of molecular modeling in relation to projects in rational drug design [Dolata,Leach and Prout,1987; Wippke and Hahn, 1988] For macromolecules, artificial intelligence tools have also been used extensively in the computer-assisted classification of structural subunits, an essential precursor to structure prediction. Numerous studies of protein structure classification and prediction, aimed at various levels of the protein structural hierarchy, have been reported (e.g., [Blundell, Sibanda, Sternberg and Thornton,1987; Clark,Barton and Rawlings,1990; Hunter and States,1991; Rawlings, Taylor, Nyakairu,Fox and Sternberg, 1985; Rooman and Wodak,1988]). In addition, promising work in application of artificial inteligence methods to the interpretation of NMR spectra of macromolecules has begun (e.g. Edwards, *et al*, this volume).

The use of artificial intelligence techniques to assist crystal structure determination, particularly in the interpretation of electron density maps, was suggested early on by Feigenbaum, Engelmore and Johnson [1977] and pursued in the CRYSALIS project [Terry,1983]. This project has not yet resulted, however, in a fully implemented and distributed system. More recently, several groups (e.g. [Finzel et al., 1990, Jones et al., 1991, and Holn and Sander, 1991] have reported the use of highly efficient algorithms for the automated interpretation of medium to high resolution electron density maps using templates derived from the Protein Data Bank [Bernstein et al.,1977] Our project, however, is concerned with the full image reconstruction process and, in particular, the ab initio phasing of diffraction data. Primarily it is the low to medium resolution region of the image reconstruction problem that is being addressed here. Clearly, our approach can draw from the many important results mentioned above.

7 Conclusion

The knowledge-based system described in this chapter offers a comprehensive approach to crystal structure determination, which accommodates a variety of phasing tools and takes advantage of the structural knowledge and experience already accumulated in the crystallographic databases. Intrinsic to our approach is the use of imagery to represent and reason about the structure of a crystal. Artificial intelligence tools that capture the processes involved in mental imagery allow us to mimic the visualization techniques used by crystallographers when solving crystal structures.

The problem of structure determination is essentially reformulated as the determination of an appropriate number of sufficiently accurate phases so as to generate a fully interpretable image of the crystal. In other words, we reduce the overall problem to a search problem in phase space. The search is guided by the continual refinement of an image through the use of partial structure information. This information is generated by matching the salient

features of the developing image with anticipated structural patterns established in previous experiments.

The process of determining the structure of a crystal is likened to an iterative scene analysis which draws both from the long-term memory of structural motifs and the application of chemical and crystallographic rules. In this analysis, the molecular scene is reconstructed and interpreted in a fluid procedure which establishes a continuum between the initially uninterpreted image and the fully resolved one. The artificial intelligence infrastructure, with its machine imagery model, allows for a coherent and efficient reconstruction. Indeed, it provides a data abstraction mechanism that can be used to reason about images and, in particular, to depict and reason with relevant configurational, conformational and topological information at a symbolic level. Thus our approach builds upon the current methodology used for protein crystal structure determination by setting a framework in which reasoning tasks as well as numerical calculations can be invoked. In this integrated approach, the process of crystal structure determination becomes one of molecular scene analysis. Taken individually, such analyzes result in the recognition and understanding of a specific chemical scene. Put together, they provide insight into the three-dimensional grammar of chemistry and the rules of molecular recognition.

Acknowledgements

Financial assistance from the Natural Science and Engineering Research Council of Canada, Queen's University and the IRIS Federal Network Center of Excellence is gratefully acknowledged.

Notes

1. The depict slot is not illustrated in the frame in Figure 2 since this function is constant for all images. For a detailed description of the frame representation for imagery and the implementation of the depict function see [Papadias 1990].

2. See [Baddeley, 1986] for an overview of algorithms for two-dimensional segmentation.

References

F. H. Allen, G. Bergerhoff, and R. Sievers, *Crystallographic Databases.* IUCr, Chester, 1987.

F. H. Allen, M. J. Doyle, and R. Taylor. Automated Conformational Analysis from Crystallographic Data. 1. A Symmetry-modified Single-linkage Clustering Algorithm for 3D Pattern

Recognition. *Acta Crystallographica*, B 47: 29-40, 1991.

F. H. Allen, O. Kennard, and R. Taylor. Systematic Analysis of Structural Data as a Research Tool in Organic Chemistry. *Accounts of Chemical Research*, 16: 146-153, 1983.

Alan Baddeley. *Working Memory.* Oxford Science Publications, 1986.

D. H. Ballard and C. M. Brown. *Computer Vision.* Prentice Hall Inc,. 1982.

F. C. Bernstein, F. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr. , M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The Protein Data Bank: A Computer Archival File for Macromolecular structures. *Journal of Molecular Biology*. 112: 535-542, 1977.

N. Block, ed. *Imagery*, MIT Press, 1981.

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

G. Bricogne. A Bayesian Statistical Theory of the Phase Problem. A Multi-channel Maximum-entropy Formalism for Constructing Joint Probability Distribution Factors. *Acta Crystallographica*, A44:517-545, 1988.

D. A. Clark, G. J. Barton and C. J. Rawlings. A Knowledge-based Architecture for Protein Sequence Analysis and Structure Prediction. *Journal of Molecular Graphics*, 8:94-107, 1990.

G. R. Desiraju, *Crystal Engineering.* Elsevier, London, 1989.

D. P. Dolata, A. R. Leach and C. K. Prout. WIZARD: Artificial Intelligence in Conformational Analysis. *Journal of Computer-Aided Molecular Design*, 1:73-86, 1987.

M. C. Etter, J. C. MacDonald and J. Bernstein. Graph-set Analysis of Hydrogen-bond Patterns in Organic Crystals. *Acta Crystallographica*, B46: 256-262, 1990.

E. A. Feigenbaum, R. S. Engelmore and C. K. Johnson. A Correlation Between Crystallographic Computing and Artificial Intelligence Research. *Acta Crystallographica*, A33:13-18, 1977.

R. A. Finke. *Principles of Mental Imagery.* MIT Press, 1989.

B. C. Finzel, S. Kimatian, D. H. Ohlendorf, J. J. Wendoloski, M. Levitt and F. R. Salemme. Molecular Modeling with Substructure Libraries Derived from Known Protein Structures. In *Crystallographic and Modelling Methods in Molecular Design*, C. E. Bugg and S. E. Ealick, eds. Springer-Verlag, New York, 1990.

S. Fortier, J. I. Glasgow, and F. H. Allen. The Design of a Knowledge-based System for Crystal Structure Determination. In H. Schenk, editor, *Direct Methods of Solving Crystal Structures.* Plenum Press, London, 1991.

S. Fortier and G. D. Nigam. On the Probabilistic Theory of Isomorphous Data Sets: General Joint Distributions for the SIR, SAS, and Partial/Complete Structure Cases. *Acta Crystallographica*, A45:247-254, 1989

J. I. Glasgow. Artificial Intelligence and Imagery. In *Proceedings of Tools for Artificial Intelligence*, Washington, 1990.

J. I. Glasgow. Imagery and Classification. In *Proceedings of the 1st ASIS SIG/CR Classification Research Workshop*, Toronto, 1990

J. I. Glasgow and D. Papadias. Computational Imagery, *Cognitive Science*, In press, 1992.

J. I. Glasgow, M. A. Jenkins, C. McCrosky and H. Meijer. Expressing Parallel Algorithms in Nial. *Parallel Computing* , 11,3:46-55 1989.

N. A. B. Gray. *Computer-Assisted Structure Elucidation.* John Wiley, New York, 1986.

L. Hache. The Nial Frame Language. Master's thesis Queen's University, Kingston, 1986

H. Hauptman. The Direct Methods of X-ray Crystallography. *Science*, 233:178 - 183, 1986.

J. B. Hendrickson. The Use of Computers for Synthetic Planning. *Angewandte Chemie International Edition (English).* , 29:1286-1295, 1990.

L. Holn and C. Sander, Database Algorithm for Generating Protein Backbone and Side-chain Coordinates from a Cα Trace: Application to model building and detection of co-ordinate errors. *Journal of Molecular Biology*, 218:183-194,1991.

L. Hunter and D. J. States, Applying Bayesian Classification to Protein Structure. in *Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications*, IEEE Computer Society Press, 1991.

M. A. Jenkins and J. I. Glasgow. A Logical Basis for Nested Array Data Structures. *Programming Languages Journal* 14 (1): 35-49, 1989.

M. A. Jenkins, J. I. Glasgow, and C. McCrosky. Programming Styles in Nial. *IEEE Software.* 86:46-55, January 1986.

T. A. Jones, J-Y. Zou, S. W. Cowan and M. Kjeldgaard. Improved Methods for Building Protein Models in Electron-density Maps and the Location of Errors in Those Models. *Acta Crystallographica*, A47:110-119, 1991.

S. M. Kosslyn. *Image and Mind.* Harvard University Press, 1980.

J. H. Larkin and H. A. Simon. Why a Diagram Is (Sometimes) Worth Ten Thousand Words. *Cognitive Science 11*: 65-99, 1987.

S. Lewis. Pattern Matching through Imagery. Master's thesis, Queen's University, Kingston, 1990.

N. MacKenzie. *Dreams and Dreaming.* Aldus Books, London, 1965.

D. Marr and H. K. Nishihara. Representation and Recognition of the Spatial Organization of Three-dimensional Shapes. In *Proc. of the Royal Society of London*, B200: 269-294, 1978.

D. Marr. *Vision.* W. H. Freeman and Company, San Francisco, 1982.

T. More. Notes on the Diagrams, Logic and Operations of Array Theory. In Bjorke and Franksen, editors, *Structures and Operations in Engineering and Management Systems.* Tapir Pub. , Norway, 1981.

D. Papadias. A Knowledge Representation Scheme for Imagery. Master's thesis, Queen's University, Kingston, 1990.

L. Pauling. *The Nature of the Chemical Bond.* Cornell University Press, Ithaca, 1939.

A. P. Pentland. Perceptual Organization and Representation of Natural Form. *Artificial Intelligence, 28* :295-331, 1986.

Z. W. Pylyshyn. The Imagery Debate: Analog Media Versus Tacit Knowledge. In N. Block, editor, *Imagery,* 151-206. MIT Press, 1981.

C. J. Rawlings, W. R. Taylor, J. Nyakairu, J. Fox and M. J. E. Sternberg. Reasoning about Protein Topology Using the Logic Programming Language PROLOG. *Journal of Molecular Graphics*, 3:151-157,1985.

M. J. Rooman and S. J. Wodak. Identification of Predictive Sequence Motifs Limited by Protein Structure Data Base Size. *Nature*, 335:45-49 1988.

M. G. Rossman. The Molecular Replacement Method. *Acta Crystallographica* A46:73-82, 1990.

R. S. Rowland, F. H. Allen, W . M. Carson, and C. E. Bugg. Preferred Interaction Patterns from Crystallographic Databases, In S. E. Ealick and C. E. Bugg, editors, *Crystallographic and Modeling Methods in Molecular Design.* Springer, New York, 1990.

L. Sawyer and M. N. G. James. Carboxyl-carboxylate Interactions in Proteins. *Nature*, 295:79-80, 1982.

R. N. Shepard and J. Metzler. Mental Rotation of Three-dimensional Objects. *Science*, 171:701-703, 1971.

R. Taylor and O. Kennard. Hydrogen-bond Geometry in Organic Crystals. *Accounts of Chemical Research*, 17:320-326, 1984.

A. Terry. The CRYSALIS Project: Hierarchical Control of Production Systems. Technical Report HPP-83-19, Stanford University, Palo Alto, CA, 1983.

J. D. Watson. *The Double Helix.* Wiley, New York, 1968.

W. T. Wippke and M. A. Hahn. AIMB: Analogy and Intelligence in Model Building. *Tetrahedron Computer Methodology*, 1:141-153, 1988.