Acta Cryst. (1993). D49, 168–178

Molecular Scene Analysis: the Integration of Direct-Methods and Artificial-Intelligence Strategies for Solving Protein Crystal Structures

BY SUZANNE FORTIER AND IAN CASTLEDEN
Department of Chemistry, Queen's University, Kingston, Canada K7L 3N6

JANICE GLASGOW, DARRELL CONKLIN, CHRISTOPHER WALMSLEY AND LAURENCE LEHERTE
Department of Computing and Information Science, Queen's University, Kingston, Canada K7L 3N6

AND FRANK H. ALLEN
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

(*Received* 12 *June* 1992; *accepted* 26 *August* 1992)

Abstract

A knowledge-based approach to crystal structure determination is presented. The approach integrates direct-methods and artificial-intelligence strategies to rephrase the structure determination process as an exercise in *scene analysis*. A general joint probability distribution framework, which allows the incorporation of isomorphous replacement, anomalous scattering and *a priori* structural information, forms the basis of the direct-methods strategies. The accumulated knowledge on crystal and molecular structures is exploited through the use of artificial-intelligence strategies, which include techniques of knowledge representation, search and machine learning.

1. Introduction

Traditional direct methods explore the phase space and evaluate phasing paths by using only very general chemical constraints - non-negativity of the electron-density distribution and atomicity - and the constraints imposed by the amplitude data. While these constraints have proven sufficiently limiting for applications to small molecules, they may not be adequate for more complex structures. The introduction of additional constraints, particularly in the form of partial structure information, has often proven crucial to the determination of small-molecule crystal structures whose complexity goes beyond that normally tackled by direct methods, *e.g.* leu-enkephalin (Karle, Karle, Mastropaolo, Camerman & Camerman, 1983) and gramicidin A (Langs, 1988). Thus, in the effort to expand the applicability of direct methods to more and more complex structures, the traditional totally data-driven and universal approach may have to be traded in for one that is context driven and that can take advantage of all available and useful information. Recent results in the theory and applications of direct methods show that such an approach can now be realized. Specifically, the development of a general joint probability distribution formalism which can integrate information from various sources, including isomorphous replacement, anomalous scattering and partial structure information, provides the framework needed for a flexible and context-driven phasing strategy.

A large amount of structural information is available, if not usually exploited, at the outset of a structure determination exercise (Allen, Bergerhoff & Sievers, 1987). Its systematic use is not trivial, in part because the information is not available in the form of synthesized knowledge but rather in the form of coordinate data, defined in terms of the external reference frame of a specific unit cell. The transformation of data to knowledge and databases to knowledge bases must first be accomplished if the vast amount of available crystallographic results is to be fully exploited.

By combining structural knowledge with the direct-methods tools, the crystallographic image reconstruction exercise can be reformulated as an information processing task and resolved through the more comprehensive approach of *scene analysis* (Duda & Hart, 1973), as earlier envisaged by Feigenbaum, Engelmore & Johnson (1977). The concept of scene analysis has been used in the context of machine vision to refer to the set of processes associated with the reconstruction, classification and understanding of complex images. Such analyses rely on the availability of *a priori* domain information, both in the form of templates and in the form of rules and heuristics, to locate and identify features in a scene and to provide for a full interpretation of the scene. By analogy, we use the phrase *molecular scene analysis* to refer to the processes associated with the reconstruction and interpretation of crystal and molecular structures.

We are currently designing and implementing a knowledge-based system for molecular scene analysis. This system incorporates direct-methods probabilistic strategies, the experience accumulated in the crystallographic databases, and knowledge-representation and reasoning techniques from artificial intelligence. In the system the process of determining the structure of a crystal

0907-4449/93/010168-11$06.00

is likened to an iterative and hierarchical scene analysis in which the search is initiated by the mathematical tools of direct methods while it is guided by pattern-recognition techniques, and rules and heuristics derived from chemistry and crystallography. Since the project is still at the design and prototyping stage, this paper does not describe a final product but rather the strategies and tools that have been selected and built to form the core of the knowledge-based system. In §2 of the paper the direct-methods strategies are presented and in §3 artificial-intelligence strategies are presented. The articulation between direct methods and artificial intelligence, and its implementation as a knowledge-based system for molecular scene analysis, is described in §4.

2. Direct-methods strategies

The method of joint probability distribution (j.p.d.) forms the basis of our direct-methods strategies for molecular scene analysis. In recent years a large amount of effort has been directed at expanding the applicability of direct methods to macromolecular structures. In particular, j.p.d.'s have been derived to integrate the techniques of direct methods with those of isomorphous replacement (*e.g.* Hauptman, 1982*a*; Fortier, Weeks & Hauptman, 1984) and anomalous scattering (*e.g.* Hauptman, 1982*b*; Giacovazzo, 1983). These results, which are reviewed in Fortier (1991), have led to the formulation of a general j.p.d. framework for application to macromolecular structures.

2.1. *General joint probability distribution framework*

It was recently shown that the j.p.d.'s for the isomorphous replacement, anomalous scattering and partial/complete structure cases are completely isomorphous: that is, they have the same functional form and differ only in individual expressions of the atomic scattering factors (Fortier & Nigam, 1989). Consequently, it is not necessary to have specialized formulae for each and every case. Instead, general j.p.d.'s can be formulated and used. In addition, it is an easy task to translate a distribution derived for a specific case into more general terms. Thus, much of the theoretical foundation available already for either the isomorphous replacement, anomalous scattering or partial/complete structure cases may be reformulated so that it can be applied to any case of interest. Distributions for a pair (Hauptman, 1982*a*) and triplet (Fortier, Weeks & Hauptman, 1984) of isomorphous structures have been fully derived, following the methods described by Karle & Hauptman (1958) and Hauptman (1982*a*). From these it is now possible to infer j.p.d.'s for any number of isomorphous data sets. Similarly j.p.d.'s have been derived for three- (Cochran, 1955; Hauptman, 1976), four- (Hauptman, 1975), five- (Fortier & Hauptman, 1977*a*) and six-phase structure invariants (Fortier & Hauptman, 1977*b*). Again these results allow distributions for *n*-phase invariants to be inferred.

Several authors have already presented procedures for a more general derivation of j.p.d.'s of structure factors. We note, in particular, the procedures described by Peschar & Schenk (1987) and Castleden (1987) for the generation of *n*-phase invariants, and the method proposed by Peschar & Schenk (1991) for the derivation of triplet invariant distributions for any number of isomorphous structure factors, which incorporates isomorphous replacement and anomalous scattering in a unified manner. It has also been shown that the j.p.d.'s are easily derived from a maximum-entropy approach (Bricogne, 1984; Bryan, 1988). To summarize, we now have at our disposal sufficient information to build j.p.d.'s of *n*-phase invariants for any number of isomorphous data sets. These distributions allow us to integrate information arising from a variety of sources, such as anomalous scattering, isomorphous replacement and *a priori* structural information. Furthermore, it is also possible to rely on the use of computer algorithms for the generation of these distributions. We have used programs to generate three-phase invariant distributions for the case of two and three isomorphous data sets. Peschar & Schenk (1987, 1991) have reported computer-aided derivations of j.p.d.'s for *n*-phase invariants and for *n*-isomorphous data sets. We can thus envisage a framework within which phasing tools, tailored to specific information contexts, are generated dynamically as needed. Such a framework would permit an opportunistic and highly flexible approach to phasing and to structure determination.

2.2. *Global aspects of direct methods: a connection with structure refinement*

The goal of a structure determination exercise can be summarized as the specification of the vector of atomic positions $\tilde{\mathbf{r}} \equiv (\mathbf{r}_1 \ldots \mathbf{r}_N)$ labelled with attributes of atomic type, thermal motion parameters and population. This goal may be attained through the use of structure factors $\tilde{\mathbf{F}}^\alpha \equiv \tilde{F}^\alpha \exp(i\tilde{\varphi}^\alpha) = \{F_{\mathbf{h}}^\alpha \exp(i\varphi_{\mathbf{h}}^\alpha)|\mathbf{h}\}$ (where $\alpha = 1 \ldots M$) derived from various scattering experiments, including isomorphous replacement and anomalous scattering, and *a priori* structural knowledge. The global probability distribution P of these values can be written, following Gillespie (1983), as:

$$P(\tilde{\mathbf{r}}, \tilde{F}^\alpha, \tilde{\varphi}^\alpha | \tilde{f}^\alpha) = \prod_{\alpha=1}^{M} \prod_{\mathbf{h}} \delta[\mathbf{F}_{\mathbf{h}}^\alpha - \mathbf{F}_{\mathbf{h}}^{\alpha,\mathrm{cal}}(\tilde{\mathbf{r}})] p_{\mathrm{prior}}(\tilde{\mathbf{r}}) \quad (2.1)$$

where it is assumed that the measurement of each reflection from each experiment is independent (in a probabilistic sense). The symbol δ is used to indicate a delta function and $\mathbf{F}_{\mathbf{h}}^{\alpha,\mathrm{cal}}(\tilde{\mathbf{r}})$ is the calculated structure factor. The expression, $p_{\mathrm{prior}}(\tilde{\mathbf{r}})$, is a prior probability distribution for the atomic positions and is set to a constant value when the atoms are assumed to be uniformly and independently distributed in the asymmetric unit. Isomorphism is enforced by the fact that the atomic positions are the same for each isomorphous data set α. The atomic scattering factors, $\tilde{f}^{\alpha}$'s, which are *assumed known*, serve to discriminate between data sets.

At the beginning of the structure determination process, the only information available may be the sets of structure-factor magnitudes derived from the diffraction experiments. It is then customary, following the approach of Hauptman & Karle (1953), to eliminate the atomic positions from (2.1) by integration: $P(\tilde{\mathbf{r}},\tilde{F},\tilde{\varphi}|\tilde{\mathbf{f}}) \rightarrow P_o(\tilde{F},\tilde{\varphi}|\tilde{\mathbf{f}})$. This is the standard structure-factor j.p.d. Here, we would like to give a simple prescription for both generating and combining j.p.d.'s. Suppose there are N scattering groups each with a number density $\rho_\mu(\mathbf{x})$, where $\mu = 1\ldots N$, and let their Fourier transform be denoted by $\mathbf{U}^{\mu}_{\mathbf{h}}$. The $\mathbf{U}^{\mu}_{\mathbf{h}}$'s are formally different from, but numerically similar to, the unitary structure factors, $\mathbf{U}_{\mathbf{h}}$, and are related to the structure factors though the relationship:

$$\mathbf{F}^{\alpha}_{\mathbf{h}} = \sum_{\mu=1}^{N} f^{\alpha}_{\mu}(\mathbf{h})\mathbf{U}^{\mu}_{\mathbf{h}} \qquad (2.2)$$

which can be written in matrix form as $\tilde{\mathbf{F}} = \tilde{\mathbf{f}}\tilde{\mathbf{U}}$. If each of the scattering groups $\rho_\mu(\mathbf{x})$ is distributed independently of the others, the probability distribution is (Castleden, 1992)

$$P(\rho_1\ldots\rho_N) \simeq \exp\left[-\sum_{\mu=1}^{N}\int_v \rho_\mu(\mathbf{x})\ln\rho_\mu(\mathbf{x})\mathrm{d}^3\mathbf{x}\right]. \qquad (2.3)$$

What is required then is that the unknown $\tilde{\mathbf{U}}$ values be re-expressed in terms of the (partially) known $\tilde{\mathbf{F}}$ and thus that the non-square matrix $\tilde{\mathbf{f}}$ be inverted. This can be carried out simply by using the generalized inverse $\tilde{\mathbf{f}}^{\#} \equiv \tilde{\mathbf{f}}^{T}(\tilde{\mathbf{f}}\tilde{\mathbf{f}}^{T})^{-1}$ so that $\tilde{\mathbf{U}} = \tilde{\mathbf{f}}^{\#}\tilde{\mathbf{F}}$. Hence, in principle, the probability distribution (2.3) can be expressed in terms of the structure-factor magnitudes and phases. In practice, the $\rho\ln\rho$ terms inside the integral sign can be expanded as a power series in the density: $\rho\ln\rho \simeq a\rho + b\rho^2 + c\rho^3$. When integrated, each term ρ^n is equal to a sum over invariants of order n. Note that an inspection of the generalized inverse shows that the separate experiments should be as orthogonal as possible, that is to say that the form factors should be as different as possible. The structure-factor magnitudes $\tilde{\mathbf{F}}^{\mathrm{obs}}$ can be held fixed at their measured values so that the distribution becomes essentially a function of the phases $\tilde{\varphi}$ which can be maximized with the tangent formula (a local steepest ascent method for phases modulo 2π). This method has a remarkable radius of convergence, especially since its complement, the least-squares method, suffers so badly. Structures of less than 100 atoms are routinely solved starting from random phases, despite the fact that these phases usually imply regions of negative electron density.

At the end of the structure determination, when a reasonable model of the structure is available, it is usually the phases that are treated as unknown and integrated out of distribution (2.1). Consider the case of a single experiment $M = 1$. Because the magnitudes are not known precisely, the δ functions in (2.1) are approximated by Gaussian functions:

$$\delta(x) \rightarrow \exp(-x^2/\sigma^2)/(\pi\sigma^2)^{1/2}$$

where σ is chosen to reflect the error in the measurement of $F_{\mathbf{h}} = F^{\mathrm{obs}}_{\mathbf{h}}$. (A more elegant argument can be made by an application of Bayes' rules). Integrating out the phases $\tilde{\varphi}$ yields (Gradshteyn & Ryzhik, 1980, integral 3.339):

$$P(\tilde{\mathbf{r}}|F^{\mathrm{obs}}) = \prod_{\mathbf{h}} I_o[2|F^{\mathrm{obs}}_{\mathbf{h}}F^{\mathrm{cal}}_{\mathbf{h}}(\tilde{\mathbf{r}})|/\sigma^2_{\mathbf{h}}]\exp\{-[|F^{\mathrm{obs}}_{\mathbf{h}}|^2 + |F^{\mathrm{cal}}_{\mathbf{h}}(\tilde{\mathbf{r}})|^2]/\sigma^2_{\mathbf{h}}\}p_{\mathrm{prior}}(\tilde{\mathbf{r}}). \qquad (2.4)$$

If the $\sigma_{\mathbf{h}} \rightarrow 0$, the modified Bessel functions can be replaced by their asymptotic forms $I_o(x) \rightarrow \exp(x)/(2\pi x)^{1/2}$ to yield:

$$P(\tilde{\mathbf{r}}|F^{\mathrm{obs}}) = \prod_{\mathbf{h}} [4\pi|F^{\mathrm{obs}}_{\mathbf{h}}F^{\mathrm{cal}}_{\mathbf{h}}(\tilde{\mathbf{r}})|/\sigma^2_{\mathbf{h}}]^{-1/2}\exp\{-\sum_{\mathbf{h}}[|F^{\mathrm{obs}}_{\mathbf{h}}| - |F^{\mathrm{cal}}_{\mathbf{h}}(\tilde{\mathbf{r}})|]^2/\sigma^2_{\mathbf{h}}\}p_{\mathrm{prior}}(\tilde{\mathbf{r}}). \qquad (2.5)$$

This procedure is easily generalized to the case of $M > 1$. The negative of the argument of the exponential in (2.5) is a chi-square function, the usual function minimized in final structure refinement. Its minimum occurs when $|F^{\mathrm{obs}}| = |F^{\mathrm{cal}}|$ but the probability (2.4) is maximized when $x = I_1(2x|F^{\mathrm{obs}}|^2/\sigma^2)I_o(2x|F^{\mathrm{obs}}|^2/\sigma^2)$, where $x = |F^{\mathrm{cal}}|/|F^{\mathrm{obs}}|$, *i.e.* when the calculated structure factors are uniformly less than the observed values and are dependent on the error estimate of the measurements.

In many crystal structure determination exercises, and particularly in those of protein crystal structures, there are several intermediate steps between the initial phasing process and the final refinement step. The structural information that is gained in these intermediate steps can be recycled to improve the phase estimates of the probability distribution (Main, 1976). From our point of view, as more and more of the structure is found, one should pass from a distribution such as (2.3) to one such as (2.4). As will be shown for the case of a known and positioned fragment, the resulting probability distribution bears a simple relationship to the original j.p.d. P_o.

Assume that in a certain region of space there is a known fragment of the structure with density $\tau(\mathbf{r})$ and let $\tau_{\mathbf{h}} = \mathrm{FT}[\tau(\mathbf{r})]$ be its Fourier transform. The atoms $\mu = 1\ldots N$, forming the unknown part of the structure must be, in some way, excluded from this region. Let the excluded region be defined by the indicative function $\chi(\mathbf{r})$ and let $\chi^{\mu}(\mathbf{r}) = \int_v \chi(\mathbf{r}')\rho_\mu(\mathbf{r} - \mathbf{r}')\mathrm{d}\mathbf{r}'$ be functions 'blurred' by the spatial extent of each atom. If $\chi_{\mathbf{h}} = \mathrm{FT}[\chi(\mathbf{r})]$ then by the convolution theorem $\chi^{\mu}_{\mathbf{h}} = \chi_{\mathbf{h}}f_\mu(\mathbf{h})$. We can thus define the prior probability as $p_{\mathrm{prior}}(\tilde{\mathbf{r}}) \propto \exp[\sum_{\mu=1}^{N}\chi^{\mu}(\mathbf{r}_\mu)]$ to down-weight the occurrence of an unknown atom within the excluded region. Calculating the Laplace transform of (2.1) ($M = 1$) with respect to the $\mathbf{F}_{\mathbf{h}}$ gives $\exp[\tilde{\lambda}\tilde{\mathbf{F}}^{\mathrm{cal}}(\tilde{\mathbf{r}}) + \sum_{\mu=1}^{N}\chi^{\mu}(\mathbf{r}_\mu)]$. Integrating out the atomic positions $\tilde{\mathbf{r}}$ will give the characteristic function $C_\chi(\tilde{\lambda}) = C_o(\tilde{\lambda} - \tilde{\chi})$.

The subscript *o* denotes the characteristic function for the distribution with no excluded region ($\chi \equiv 0$) and the equality is an obvious corollary of the definition of $\chi^{\mu}_{\mathbf{h}}$. The inverse Laplace transform performed with a change of variables $\widetilde{\lambda}' = \widetilde{\lambda} - \widetilde{\chi}$ gives:

$$P(\widetilde{\mathbf{F}} - \tau) = \{\exp[\widetilde{\chi}(\widetilde{\mathbf{F}} - \tau)]/Z(\widetilde{\chi})\}P_o(\widetilde{\mathbf{F}} - \widetilde{\tau}) \qquad (2.6)$$

where **Z** is a normalization term:

$$Z(\chi) = V^{-N}\int_v p_{\text{prior}}(\widetilde{\mathbf{r}})\,\mathrm{d}\widetilde{\mathbf{r}} = \prod_{\mu=1}^{N}\int_v \exp[\chi^{\mu}(\widetilde{\mathbf{r}}_{\mu})]\,\mathrm{d}\widetilde{\mathbf{r}}_{\mu}/V. \quad (2.7)$$

This relationship has been used successfully by Beurskens *et al.* (1981) in the *DIRDIF* program with the approximation that $\chi \equiv 0$ and that the magnitude $|\mathbf{F}_{\mathbf{h}} - \tau_{\mathbf{h}}\exp(-i\varphi_{\mathbf{h}})|$ is independent of $\varphi_{\mathbf{h}}$. The latter condition is approximately true if either $\tau_{\mathbf{h}}$ or $\mathbf{F}_{\mathbf{h}}$ is small or if the initial phase estimate is good so that $\Delta\widetilde{\varphi}$ remains small during tangent refinement. Note that, using Bayes' rule, the known part of the structure τ can be considered as a random variable. In this case, however, careful attention must be paid to the variation of $Z(\chi)$ which is effectively a function of τ.

Equations (2.1)-(2.7) make clear the connection between the initial direct-methods phase determination and the final structure refinement and show that they are two different aspects of the same underlying distribution. The structure-factor j.p.d. is used when the atomic positions are completely unknown. This lack of knowledge is reflected by the fact that the atomic positions are integrated out. Once they have been found then our lack of knowledge about the structure-factor phases is evoked and one refines only on the magnitudes. In between are the cases of partial structure knowledge. Greater structure-solving power may be available once these different methods are taken as two extremes on a continuum and the path between initial phase determination and final chi-square refinement is 'joined' by a series of steps characterized by equation (2.6).

3. Artificial-intelligence strategies

Through the general joint-probability framework an extensive array of tools can be tailor-built to accommodate specific problems in protein crystal structure determinations. There is also a rich knowledge base from which to draw for the recovery and interpretation of protein images. Included in this knowledge base is the vast reservoir of structural data available in the crystallographic databases, as well as the rules that have so far been found to govern molecular architecture and the heuristics and methodologies that have proved useful in the reconstruction of crystal and molecular structures. We have turned to artificial-intelligence strategies to develop a coherent framework within which the available molecular knowledge can be integrated into the phasing tools.

The use of solution strategies that can draw from general and domain knowledge and that can also learn from past experience has been central to artificial-intelligence approaches to problem solving. Artificial intelligence has long been concerned with the question of how to organize and represent knowledge so as to allow for the rapid and efficient retrieval of information on which reasoning tasks can then be performed. The development of search strategies is also an important activity in the field of artificial intelligence, where it is customary to rephrase problems as search problems. Solutions can then be formulated in terms of strategies that are used to represent and explore the search space and in terms of functions that are used to evaluate search states. Finally, much effort has gone into the area of machine learning to develop techniques for the classification and extension of knowledge. The results of research in knowledge representation, search strategies and machine learning can contribute to the design and development of a scene-analysis approach to crystal structure determination.

3.1. *Knowledge representation*

In molecular scene analysis we are primarily concerned with questions pertaining to shape and spatial relationships whose answers rely, in part, on the efficient recall and analysis of previously determined molecular scenes, and on the application of the rules and heuristics of chemistry. We have, therefore, developed a knowledge representation scheme that consists of three interrelated representations: a descriptive representation, which serves as a knowledge base of molecular scenes, and two depictive representations which are used to carry out the spatial and visual analysis of an image (Glasgow, Fortier & Allen, 1992).

The molecular knowledge base is being implemented as a frame system (Minsky, 1975) using the Nial frame language (Hache, 1986). Information in the system is organized according to both the structural and conceptual hierarchies of molecular structures. As shown in Fig. 1, the structural hierarchy separates the information according to the building block model of molecular structures while the conceptual hierarchy organizes information according to the classes, subclasses and instances of the structural building blocks. Individual frames in the system are used to cluster information on a given structural element or on a class of structural elements. As illustrated in Fig. 2, an individual frame is defined as a simple data object which organizes the information into slot-value pairs. It is the high level of flexibility in the type of values a slot can take, *e.g.* numeric and literal data, pointers to other frames, structured data, attached procedures *etc.*, that makes a frame a versatile and all-encompassing knowledge representation tool. It is also possible to attach constraint specifications to a slot-value pair. For example, values to be added can be constrained to a specified data type or to a given numerical range. In addition, conditions such as 'if-needed', 'if-added' or 'if-removed' can be associated with a slot to limit or trigger the execution of procedures. In the knowledge base, individual frames are linked to one another through semantic links. Of particular importance are the PART_OF and AKO (a kind of) links which establish the dual structural and conceptual hierarchies of the system and the inheritance pathways. Properties or values can thus be stored at the most general level of the conceptual hierarchy and be inherited by the more specific levels. A frame system can generally be viewed as a taxonomy of linked concepts with the knowledge, both descriptive and procedural, organized around each concept, similar to the object-oriented database model. What further characterizes a frame system is its structural flexibility. In particular, frames in the system can be instantiated either in a static or dynamic mode. In the latter case, both a frame and its links to the rest of the frame system can be created dynamically, thus providing an environment capable of continued modification, expansion and enhancement.

Algorithms for creating frame structures for entries retrieved from the Cambridge Structural Database (Allen *et al.*, 1991) and the Protein Data Bank (Bernstein *et al.*, 1977) have been implemented. A system for browsing through large knowledge bases of frames is also being developed (Martin, Hung & Walmsley, 1992) in parallel with the design and implementation of the molecular knowledge base. In this browsing system, the knowledge base is represented as a graph consisting of nodes and arcs, as illustrated in Fig. 3. The nodes are used to represent frames or other data values while the arcs represent semantic links or slots. An initial query serves to define a starting node location in the knowledge base. The node and its immediate and/or extended neighborhood of arcs and nodes are displayed and the knowledge base can then be further explored by following the arcs to connected nodes and focusing on new nodes. Focusing on a new frame node, for example, brings that node to the centre of the graph, displaying its extended neighborhood. Focusing on a procedure value triggers the execution of the procedure, the results of which are then displayed in a new window. The system supports both textual and graphical capabilities for querying and viewing the knowledge base. Complex graphical queries are easily constructed with editing capabilities that allow for nodes and arcs to be selected, deleted, copied, pasted, moved and labelled. The textual query facility provides additional flexibility by allowing queries to be embedded into programs.

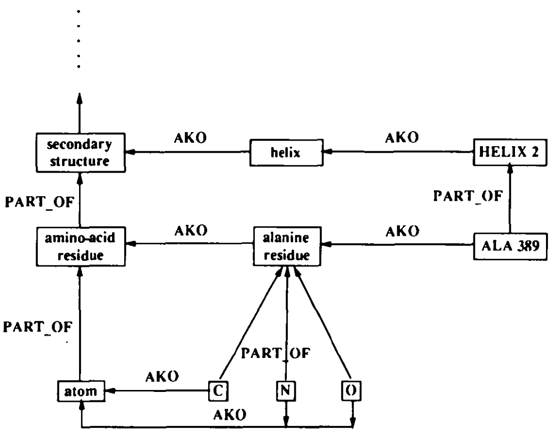
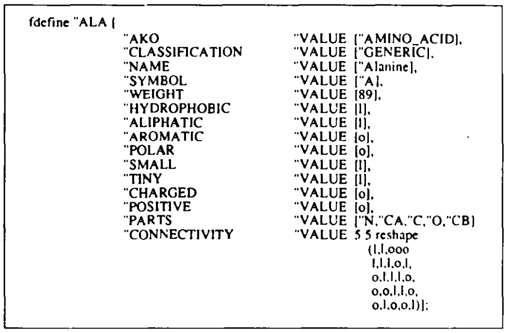


Fig. 1. Structural and conceptual hierarchies in the molecular knowledge base.

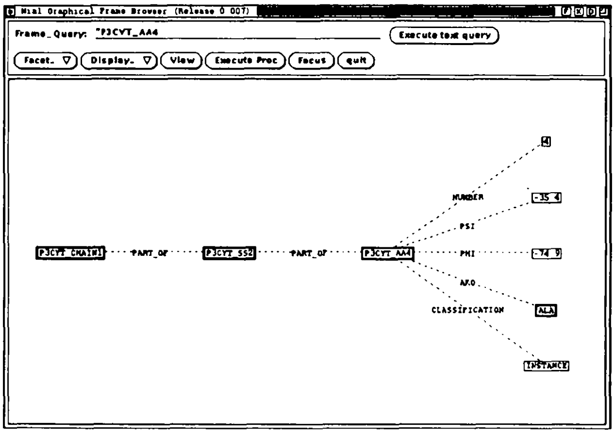


```
fdefine "ALA [
            "AKO                  "VALUE ["AMINO_ACID],
            "CLASSIFICATION       "VALUE ["GENERIC],
            "NAME                 "VALUE ["Alanine],
            "SYMBOL               "VALUE ["A],
            "WEIGHT               "VALUE [89],
            "HYDROPHOBIC          "VALUE [l],
            "ALIPHATIC            "VALUE [l],
            "AROMATIC             "VALUE [o],
            "POLAR                "VALUE [o],
            "SMALL                "VALUE [l],
            "TINY                 "VALUE [l],
            "CHARGED              "VALUE [o],
            "POSITIVE             "VALUE [o],
            "PARTS                "VALUE ["N,"CA,"C,"O,"CB]
            "CONNECTIVITY         "VALUE 5 5 reshape
                                          (l,l,ooo
                                           l,l,l,o,l,
                                           o,l,l,l,o,
                                           o,o,l,l,o,
                                           o,l,o,o,l)];
```

Fig. 2. Frame data structure for Ala residue.

An important part of protein crystal structure determination is the direct inspection of electron-density maps so as to determine the location and the identity of structural features. It is now recognized that two separate and distinct modes of reasoning, the visual and spatial modes, are used in the high-level processes associated with image interpretation (Glasgow & Papadias, 1992) and we believe that both modes should be incorporated into automated map-interpretation algorithms. Our knowledge-representation scheme thus includes two additional components: a visual and a spatial representation. These are viewed as working-memory representations since they are normally generated, as needed, from either the diffraction data or from the descriptive information retrieved from the knowledge base.

Occupancy arrays are the usual way of depicting electron-density maps. These three-dimensional arrays make explicit the shape and relative-size information and can thus be used for visual reasoning. It is usually necessary, however, to segment the arrays into parts or blobs



Frame_Query: "P3CVT_AA4

Fig. 3. Graphical frame browser.

prior to recognition. In addition, the information contained in occupancy arrays is often too detailed to allow for extensive computational comparison. Therefore, it is essential to translate the occupancy-array representation into a simpler one that can capture the relevant shape information at varying resolution levels and discard unnecessary and distracting details. For example, Greer (1974) proposed a skeleton representation to aid in automated interpretation of protein electron-density maps. A simplified visual representation is also provided by the topological approach, which depicts the shape properties of the electron-density distribution through the identification and location of its critical points (Smith, Price & Absar, 1977; Bader, 1992). It not only provides information on the shape properties of the electron-density distribution but also serves as a segmentation tool. Such an approach has been proposed previously for the interpretation of protein electron-density maps (Johnson, 1976, 1977; Grosse, 1980). It has been implemented in the program *ORCRIT* where, in particular, the canonical features of the electron-density distribution are captured through the representation of the network of critical points (Johnson, 1976, 1977). We are currently using *ORCRIT* to identify and classify motifs of peak, pass, pale and pit critical points in secondary structures and in amino-acid residues at varying resolution levels so as to construct topological templates for use in pattern recognition (Leherte, Fortier & Glasgow, 1992).

To allow for spatial reasoning, we require a representation that explicitly denotes the spatial relationships among the parts of a scene, by analogy to the mental maps created by humans. Thus our scheme includes a spatial representation which is implemented in the form of embedded symbolic arrays. As illustrated in Fig. 4, the meaningful parts of a scene are denoted as components of an indexed array while the hierarchical structure of the scene is captured through the embedded nature of the array. Symbolic arrays differ from occupancy arrays in several respects. Firstly, they provide an explicitly interpreted depiction of a molecular scene while occupancy arrays are essentially uninterpreted. Logical functions can then be used to reason with the information encoded in the symbolically denoted features and their spatial relationships. Secondly, symbolic arrays provide a simplified depiction of a scene and, in particular, one which preserves the spatial relationships but not necessarily the precise geometry. Many questions of importance in molecular scene analysis can be answered or at least filtered, through this simplified representation (*e.g.* questions pertaining to secondary and supersecondary motifs in proteins, hydrogen-bonding motifs and configurational assignments). The information needed to answer such questions could also be encoded as propositions. The advantage of symbolic arrays over propositional representations lies in their succinct and holistic encoding as well as their provision for updating and change. These advantages translate into computational efficiency in querying functions. The scheme for knowledge representation presented here is further elaborated elsewhere (Glasgow, Fortier & Allen, 1992).

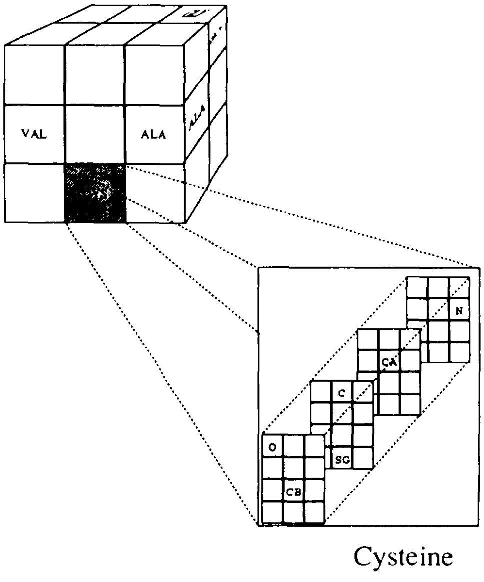


Fig. 4. Embedded symbolic arrays for the spatial representation.

3.2. *Search strategies*

Search strategies have long been central to direct-methods procedures. Indeed, the crystallographic phase problem is generally solved as a search problem. In the initial state only a handful of phases is known. In spanning the phase space, it is hoped that a final or goal state will be reached in which enough phases will have been determined with sufficient accuracy to compute an interpretable electron-density map. In this context, the multisolution approach, resulting either from permuted phases or from random starting phases, can be described as a simple generate-and-test search procedure. The morphology of the search tree is unusual, though. The tree normally has a single depth level with a large branching factor. Nevertheless, this search tree has proven highly successful and efficient for most small-molecule crystal structures. When used with more complex structures, however, it usually meets with failure for a number of reasons. Firstly, as the complexity of the problem increases, the number of possible solutions that must be explored often exhausts normal computing resources. Secondly, the reliability of the commonly used figures of merit tends to decrease as the complexity of the problem increases: a good solution may be developed but it may not be possible to recognize it. Finally, the number of false or local minima may increase with the complexity of the problem thus decreasing the chance of finding the global minimum. Alternative strategies have been used to circumvent these problems. For example, the simulated-annealing search strategy has recently been proposed as a way of increasing the chance of finding the global minimum of the minimizing function, rather than local ones (Sheldrick, 1990). This approach has already proven successful in a number of cases which had previously resisted the more standard direct-methods search procedures. To reduce computing time, *SHELX*86 (Sheldrick, 1990) introduced a two-stage phase refinement technique. Similarly, *QTAN* (Langs & DeTitta, 1975) explores the phase space using a multilevel search tree. The inability of the commonly used figures of merit to prune partially developed phase sets has been, however, a limiting factor. This has emphasized the need for a more discriminating figure of merit, a problem addressed by the recently proposed likelihood criterion (Gilmore, Bricogne & Bannister, 1990). In a number of small-molecule crystal structures, likelihood has proven extremely useful for picking the best solutions among partially developed phase sets containing only a handful of phases. It has also been shown capable of selecting the best phase sets when applied to 2 Å data from the small protein avian pancreatic polypeptide (APP) (Gilmore, Henderson & Bricogne, 1991).

With the possibility of embedding direct methods into a multilevel search tree, crystal structure determination can be formulated as an iterative and hierarchical image reconstruction exercise: the phases are determined in shells of increasing resolution yielding images that reveal structural elements of corresponding resolutions (molecular envelope, domains, secondary structures, residues). This approach not only takes advantage of the structural organization of protein but also allows for the active use of partial structure information in the phasing process: any identified structural elements can be integrated into the phasing tools *via* the general joint probability distribution framework. Such an approach requires, however, the evaluation and interpretation of a possibly large number of electron-density maps which, to be realizable, calls for substantial computer assistance.

Image understanding can itself be formulated as a search problem and, specifically, as a constraint satisfaction problem. In a crystal structure determination, features of an electron-density map are interpreted in terms of the expected chemical constitution of the crystal. In addition, the interpretation must conform to chemical and crystallographic constraints, as established from earlier experiments. Thus an approach based on constraint satisfaction can be followed that is similar to the one adopted in the knowledge-based system *PROTEAN* for elucidating protein structures from NMR data (Hayes-Roth *et al.*, 1985).

Constraint satisfaction is the process of assigning domain values to a set of specified variables such that a set of constraints is satisfied. Constraints are normally expressed as binary predicates but *n*-ary constraints are also possible. Major *et al.* (1991) have defined a constraint-satisfaction algorithm for macromolecular modelling as follows. Given the variables $X = \{x_1, x_2 \ldots x_n\}$ whose values are taken from the domains of permitted values $D = \{d_1, d_2 \ldots d_n\}$ and a set of constraints $C = (c_{p,q} \ldots | p \in \{1 \ldots n\}, q \in \{1 \ldots p-1\})$, a solution is defined in terms of the values of X, taken from its domain of values D, that satisfy all constraints in C. In molecular scene analysis, the interpretation of an electron-density map at, for example, a resolution corresponding to the residue level can be phrased as the formulation of hypotheses for the segmented features, denoted by $x_1, x_2 \ldots x_n$. The domains, d_1, $d_2 \ldots d_n$, of possible values for each feature may be restricted to subsets of amino-acid residues, according to sequence information, secondary structure information, *etc.* In addition, the hypotheses may have to satisfy a set of constraints D related to spatial and geometrical relationships among amino-acid residues. This definition of constraint satisfaction can be integrated into our knowledge-based system of frames. In the frame slot-and-filler representation, undetermined variables are denoted by unfilled slots while domains of permitted values are defined through the constraint specification facilities of frames. *N*-ary constraints, with $N > 1$, can be specified in the attached procedures that are executed on an 'if-added' basis. Both declarative and procedural constraints can be expressed at the most appropriate level of the conceptual hierarchy since constraints specified at a generic level can be inherited by subclasses or instances through the AKO links. For example, an instance of glycine can inherit constraints from the general class 'amino-acid residue' and from the subclass 'glycine'. Finally, constraints related to specific levels of the protein structural organization, *e.g.* secondary-structure architecture, atomic structure of residues, van der Waals radii of atoms, can be expressed at the appropriate level of the structural hierarchy.

Within our knowledge-representation scheme, the image reconstruction process can be defined as the transformation of an uninterpreted occupancy array into a fully interpreted symbolic array. By combining direct-methods and artificial-intelligence strategies, this goal is broken down into several subgoals which are met by the acquisition of partial structure information at the various resolution stages of the reconstruction hierarchy. At the beginning of the process, a frame is constructed for the uninterpreted image. Any available information - *e.g.* unit-cell parameters, density, primary structure - is added to the frame. The instantiation of a frame for the new scene also creates the appropriate semantic links with other frames in the system. The initial state in the search space is an uninterpreted or partially interpreted symbolic array, associated with the initial occupancy array. New states (nodes) are generated in the search tree by the application of state transformation rules which consist of formulating structural hypotheses for fragments/blobs within the current scene. This is achieved by using both the spatial and visual representations and the functions that operate on them. In particular, the spatial representation is used to analyze existing neighborhoods within the current scene so as to anticipate possible fragments based on the known context. The visual representation, on the other hand, is used to compare (pattern match) anticipated fragments with unidentified blobs based on visual features such as shape and relative size.

A crucial component of the search paradigm is an evaluation function that measures the *goodness* of partially interpreted images. This function is used in the control strategy to guide the search towards the goal of a fully interpreted image. We are presently developing such a function for the evaluation of reconstructed images of molecular scenes (Konstandinos, 1992). When evaluating these images, the important questions are whether the evolving scene fits the experimental evidence and is consistent with the constraints of chemistry and crystallography, or with similar scenes that have been previously determined, and whether the interpretation results in added knowledge. Constraints may be 'hard' or 'soft'. A state that does not satisfy a hard constraint is immediately discarded. Soft constraints, on the other hand, are used to provide evidence either for or against a state. Thus the image evaluation is achieved through the use of (1) the hard chemical and crystallographic constraints, (2) the softer constraints related to adherence to experimental evidence or to previous experience, (3) a measure of the knowledge gain, and (4) the direct-methods figures of merit. The evaluation function is used to guide the image reconstruction exercise by identifying the best phasing paths to be further developed. It is important to note that if it turns out that a bad path has been chosen, the search procedure can backtrack and follow alternative paths. Parallel exploration of the search tree is also possible.

3.3. *Learning*

Learning techniques can play an important role in the analysis, organization and compression of the vast and rapidly growing reservoir of structural data and, therefore, in the transformation of the databases into knowledge bases. Although several general concepts, rules and constraints about molecular structures have been explicitly formulated, many still remain buried within the databases. In addition, machine learning can contribute strategies for improving the performance of a knowledge-based system by indexing solution scenarios and providing a mechanism for the system to learn from its experiences. Several paradigms of machine learning have been proposed, for example the connectionist, genetic, analytic and inductive paradigms (Carbonell, 1989). In the molecular scene-analysis project, we have concentrated on the inductive approach and, in particular, the concept formation approach in which learning proceeds through the generalization, characterization and organization of a set of examples.

When reasoning about molecular structures and/or solution strategies for their reconstruction, we often proceed in one of the following ways. We may have seen several instances of a given molecular fragment and have observed that some of its structural characteristics are the same in all of the instances. We can thus form a general concept which represents and subsumes these instances and can then be used for inference. This follows the generalization-based approach to reasoning, which includes the use of general concepts, rules, constraints, *etc*. Alternatively, when analyzing a given molecular fragment we may turn not to general concepts but rather to specific instances that have a high degree of similarity with it. In this case-based approach, inference proceeds through the retrieval of similar cases, followed by comparison and adaptation (Riesbek & Schank, 1989). Clearly, both reasoning paradigms rely on some form of learning through which concepts or instances have been discovered, compared, classified or indexed. Both the generalization-based and the case-based reasoning approaches can be accommodated by our knowledge base, which includes frames for both instances and general concepts, and thus stores both individual cases and their generalizations. In addition, since the frames are linked to one another through the AKO arcs of the conceptual hierarchy, the concept frames serve as indices for the instances to which they are linked. Through these indices the subset of the most relevant cases, to be used in a case-based reasoning approach, can be retrieved. Because of the large number of individual cases, this pre-selection step is essential for computer efficiency.

The majority of the techniques that have been used to acquire knowledge from the geometrical information stored in the crystallographic databases are numerical or statistical in origin. These techniques, which have been extensively used in small-molecule applications, have recently been reviewed by Taylor & Allen (1992). Methods that are more closely allied to the artificial-intellegence approaches have also been used, particularly for the classification and prediction of protein structures (*e.g.* Hunter & States, 1991; Lathrop, Webster & Smith, 1987; Rooman & Wodak, 1988; Cohen, Abarbanel, Kuntz & Fletterick, 1986; Qian & Sejnowski, 1988; Blundell, Sibana, Sternberg & Thornton, 1987).

For many applications it is not possible to describe the parameters of interest in numerical terms. In addition, the desired result of a classification exercise is often not a set of numbers or statistics but rather a broader conceptual categorization and definition. These concerns have been addressed by the artificial-intelligence community through extensive research in machine learning and, in particular, in the areas of conceptual clustering and concept formation. Concept formation is concerned with the organization of knowledge into concept hierarchies that can then be used to explain unclassified instances (Gennari, Langley & Fisher, 1989). In general, concept formation methods are incremental, translating a stream of instances into a concept hierarchy that organizes and summarizes the instances. New concepts are discovered by the process of *generalization.*

Most work on concept formation relies on describing objects in terms of a list of attribute-value pairs. Such a representation is clearly too restrictive for the molecular domain, where spatial relations among objects must also be described. An emerging area of interest in machine learning is structured concept formation, in which structured objects - described in terms of components and their interrelationships - are formed and organized in a knowledge base (Thompson & Langley, 1991). We have designed and implemented an incremental conceptual algorithm specifically for objects or scenes composed of many parts (Conklin & Glasgow, 1992). The *I-MEM* (image-memory) system manipulates objects or scenes in terms of parts or relationships among these parts, rather than attribute-value pairs. It rests on a theory of image subsumption, based on preservation of parts relationships, and uses *part deletion* as its generalization operator. Since similarity is defined in terms of part removal, a multilevel hierarchy of concepts is obtained allowing for the comparison of the full molecular fragments and of their subfragments. Thus when a new instance is presented to the established hierarchy, the system can classify it at the appropriate level of the hierarchy, including at the subfragment levels. At the end of a run the system integrates the results automatically by creating descriptions of the acquired concepts and by establishing the subsumption links among concepts and instances. In this way, the initial collection of database entries is transformed into a compressed knowledge base. The system reads database entries translated into the frame representation and creates a knowledge base of frames as described in §3.1. *I-MEM* has been tested on examples retrieved from the small-molecule domain but is now being expanded for applications to macromolecules. A more detailed description of the *I-MEM* approach is available in Conklin & Glasgow (1992) and Conklin, Fortier, Glasgow & Allen (1992).

Because the databases are growing rapidly, the need for automated knowledge acquisition methods is also growing. It is particularly important to develop methods that can be used with little, if any, user intervention so that knowledge bases, once created, can be updated on a continued basis.

4. Knowledge-based approach to molecular scene analysis

Fig. 5 illustrates our proposed algorithm for molecular scene analysis, which consists of five independent but communicating processes. The *image-anticipation* process involves the retrieval of motifs from the *knowledge base* according to the available chemical and structural information. In the *image-enhancement and segmentation* process, the experimental electron-density map is subjected to the standard noise-reduction and density-modification routines prior to its segmentation into distinct blobs/regions that correspond to the structural features of the image. The *pattern-matching* process involves the comparison of the unidentified features, derived from the segmentation of the map, with the anticipated motifs. The possible interpretations are then analyzed for their global consistency and ranked in the *scene-analysis* process. Finally, in the *resolve* process the gained information is integrated into the direct-methods tools so as to refine and expand the phases and construct a new electron-density map.

The algorithm is applied iteratively so as to reconstruct and interpret images of progressively higher resolution. Its final goal is a fully interpreted molecular scene, at a resolution matching that of the diffraction data. One iteration through the algorithm corresponds to one level of generation in the search tree. It involves considering a particular state in the tree, generating the *offspring* of that state through the formulation of possible image interpretations and evaluating the hypothesized scenes so as to determine the next best state. The acquired partial structure information is then integrated into the phasing tools so as to refine and expand the phases and generate new images. This knowledge-based approach follows the usual steps of a crystal structure determination which is not surprising since the approach was designed to mimic the procedure normally taken by 'experts' in the field. Its novel aspect and its main contribution are in the effort to substitute many of the human interventions by computational processes.

Strategies emanating from the direct-methods and artificial-intelligence fields were presented earlier. The following is a brief summary of how these strategies contribute to the processes in the algorithm for molecular scene analysis. The image-anticipation process relies on the organization of the structural information into a knowledge base of frames. It is through the AKO links of the frame system that structural templates can be inferred or anticipated given the available chemical information. Cen-

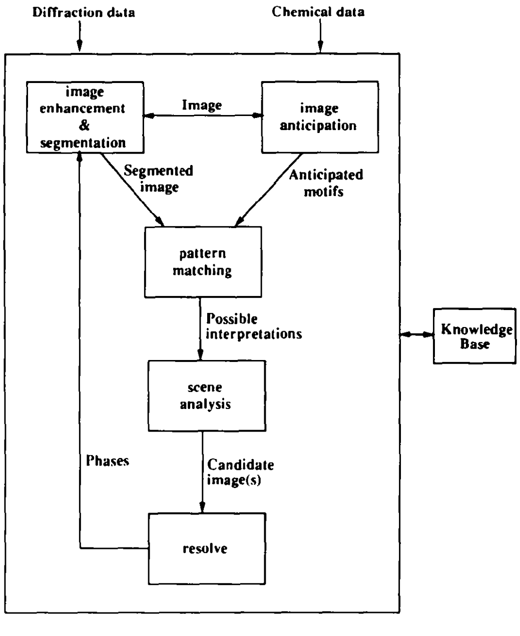


Fig. 5. Algorithm for molecular scene analysis.

tral to the creation of a knowledge base and the process of template anticipation are the learning techniques described in §3.3. Several common techniques of image processing, *e.g.* noise reduction, local averaging, ensemble averaging *etc.*, are routinely applied to protein electron-density maps to enhance their features and facilitate their interpretation. The segmentation of electron-density maps remains a difficult problem, though, particularly at low resolution. Critical point mapping techniques (§3.1) are being tested to address this problem and are an important component of the process that deals with *image enhancement and segmentation*. The iterative and hierarchical reconstruction of a molecular scene will require the analysis and interpretation of a possibly large number of electron-density maps. The algorithms in the *pattern-matching* process must, therefore, be robust and efficient. To meet these requirements, our knowledge representation scheme includes descriptive, spatial and visual components so that pattern matching can itself be formulated as a hierarchical process which proceeds from a coarse level to progressively more detailed ones. It is in the *scene-analysis* process that the combined artificial-intelligence strategies of knowledge representation, search and learning come together. Rephrasing the image-evaluation process as a search problem and, specifically, as a constraint satisfaction problem calls for the ability to (1) access a well organized and evolving knowledge base, (2) inspect the spatial and visual properties of the image and (3) prune the search tree through a robust heuristic evaluation function. Finally, the *resolve* process, which involves the resolution and reconstruction of the evolving image, relies on the direct-methods strategies described in §2.1 and §2.2 to build dynamically the phasing tools that integrate all available and useful information and expand the shells of estimated phases to a higher resolution.

A prototype system for molecular scene analysis is presently being implemented using the functional programming language Nial (Jenkins, Glasgow & McCrosky, 1986). The nested array data structure and primitive functions of Nial allow for simple manipulations of three-dimensional lattices. In addition, Nial provides, through its artificial-intelligence toolkit (Jenkins *et al.*, 1988), an array of programs and techniques that can be used to custom-build knowledge-based systems. Of particular interest for our application are the symbolic computing facilities, the Nial frame language and the ability to treat programs as data. Provisions are being made for possible reimplementation, at a later stage, of parts of the code in compiler-based or parallel languages.

5. Concluding remarks

Except for a few very simple cases, the reconstruction of three-dimensional molecular scenes is rarely accomplished solely from direct techniques. In most cases it relies extensively on the pattern-recognition and reasoning abilities (often mistaken for intuition) of an expert crystallographer. The goal of the knowledge-based approach to molecular scene analysis proposed here is to make use of the extensive amount of information now available on crystal and molecular structures so that it can guide the image-reconstruction process. By taking advantage of the complementary strengths of direct methods and artificial intelligence, it is hoped that crystal structure determination can be rephrased as an information-processing task and implemented in a coherent and comprehensive computational framework. Some of the theories, tools and approaches that are being used to build a knowledge-based system for molecular scene analysis have been described here, although we acknowledge that much work remains to be done before these translate into a fully operational system. We are currently concentrating our efforts on the construction of a knowledge base of protein crystal structures, the implementation and testing of routines for the automated interpretation of electron-density maps, the extension and application of the *I-MEM* conceptual-clustering approach, the design of a constraint-satisfaction algorithm for molecular scene analysis and, finally, the implementation of the general joint probability distribution framework. Each sub-project can independently yield useful results for the analysis of protein structures.

In this paper we have presented our vision of molecular scene analysis and, in particular, of protein crystal structure determination. It is a vision that strives to meet the delicate balance set by Barrow & Tenenbaum (1981) so as to avoid 'oversight' (not seeing things that are really present), while not suffering from 'hallucination' (seeing things that are not present at all).

Financial assistance from the Natural Sciences and Engineering Research Council of Canada in the form of operating grants to SF and JG and from NATO in the form of a collaborative research grant to JG, SF and FHA is gratefully acknowledged.

References

ALLEN, F. H., BERGERHOFF, G. & SIEVERS, R. (1987). *Crystallographic Databases.* Chester: IUCr.

ALLEN, F. H., DAVIES, J. E., GALLOY, J. J., JOHNSON, O., KENNARD, O., MACRAE, C. F., MITCHELL, E. M., MITCHELL, G. F., SMITH, J. M. & WATSON, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187-204.

BADER, R. F. W. (1992). *Atoms in Molecules.* Oxford Univ. Press.

BARROW, H. G. & TENENBAUM, J. M. (1981). *Proc. IEEE,* **69**, 572-595.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535-542.

BEURSKENS, P. T., BOSMAN, W. P., DOESBURG, H. M., GOULD, R. O., VAN DER HARK, TH. E. M., PRICK, P. A. J., NOORDIK, J. H., BEURSKENS, G. & PARTHASARATHY, V. (1981). Tech. Rep. 1981/2. Crystallography Laboratory, Toernooiveld, 6525 ED Nijmegen, The Netherlands.

BLUNDELL, T. L., SIBANA, B. L., STERNBERG, M. J. E. & THORNTON, J. M. (1987). *Nature (London),* **326**, 347-352.

BRICOGNE, G. (1984). *Acta Cryst.* **A40**, 410-445.

BRYAN, R. K. (1988). *Acta Cryst.* **A44**, 672-677.

CARBONELL, J. (1989). *Machine Learning Paradigms and Methods.* Amsterdam: Elsevier.

CASTLEDEN, I. R. (1987). *Acta Cryst.* **A43**, 384-393.

CASTLEDEN, I. R. (1992). *Acta Cryst.* **A48**, 197-209.

Cochran, W. (1955). *Acta Cryst.* **8**, 473-478.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J (1986). *Biochemistry*, **25**, 266-275.

Conklin, D., Fortier, S., Glasgow, J. I. & Allen, F. H. (1992). Proc. ML 92 Workshop Mach. Discov., Aberdeen, Scotland.

Conklin, D. & Glasgow, J. I. (1992). In *Machine Learning: Proceedings of the Ninth International Conference*, edited by D. Sleeman & P. Edwards. Los Altos: Morgan Kaufmann.

Duda, R. O. & Hart, P. E. (1973). *Pattern Recognition and Scene Analysis.* New York: John Wiley.

Feigenbaum, E. A., Engelmore, R. S. & Johnson, C. K. (1977). *Acta Cryst.* **A33**, 13-18.

Fortier, S. (1991). In *Direct Methods of Solving Crystal Structures*, edited by H. Schenk. New York: Plenum Press.

Fortier, S. & Hauptman, H. (1977*a*). *Acta Cryst.* **A33**, 572-575.

Fortier, S. & Hauptman, H. (1977*b*). *Acta Cryst.* **A33**, 694-696.

Fortier, S. & Nigam, G. D. (1989). *Acta Cryst.* **A45**, 247-254.

Fortier, S., Weeks, C. M. & Hauptman, H. (1984). *Acta Cryst.* **A40**, 646-651.

Gennari, J. H., Langley, P. & Fisher, D. (1989). *Artif. Intell.* **40**, 11-61.

Giacovazzo, C. (1983). *Acta Cryst.* **A39**, 585-592.

Gillespie, D. T. (1983). *Am. J. Phys.* **51**(6), 520-533

Gilmore, C. J., Bricogne, G. & Bannister, C. (1990). *Acta Cryst.* **A46**, 297-308.

Gilmore, C. J., Henderson, N. & Bricogne, G. (1991). *Acta Cryst.* **A47**, 842-846.

Glasgow, J. I., Fortier, S. & Allen, F. H. (1992). In preparation.

Glasgow, J. I. & Papadias, D. (1992). *Cogn. Sci.* **16**, 355-394.

Gradshteyn, I. S. & Ryzhik, I. M. (1980). *Table of Integrals, Series and Products.* New York: Academic Press.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279-301.

Grosse, E. H. (1980). PhD Thesis, Stanford Univ., Stanford, USA.

Hache, L. (1986). MSc Thesis, Queen's Univ., Kingston, Canada.

Hauptman, H. (1975). *Acta Cryst.* **A31**, 671-679.

Hauptman, H. (1976). *Acta Cryst.* **A32**, 877-882.

Hauptman, H. (1982*a*). *Acta Cryst.* **A38**, 289-294.

Hauptman, H. (1982*b*). *Acta Cryst.* **A38**, 632-641.

Hauptman, H. & Karle, J. (1953). *Solution of the Phase Problem. I. The Centrosymmetric Crystal. ACA Monograph.* Wilmington: The Letter Shop.

Hayes-Roth, B., Buchanan, B., Lichtarge, O., Hewett, M., Altman, R., Brinkley, J., Cornelius, C., Duncan, B. & Jardetzky, O. (1985). *Elucidating Protein Structure from Constraints in PROTEAN.* Tech. Rep. KSL-85-35. Stanford Univ., USA.

Hunter, L. & States, D. J. (1991). Proc. IEEE Conf. Artif. Intell. Appl., Miami, Florida, USA.

Jenkins, M. A., Glasgow, J. I., Blevis, E., Chau, R., Hache, E. & Lawson, D. (1988). Proc. Avignon '88 8th Int. Workshop Expert Syst. Appl., Avignon, France.

Jenkins, M. A., Glasgow, J. I. & McCrosky, C. (1986). *IEEE Software*, **86**, 46-55.

Johnson, C. K. (1976). Proc. Am. Crystallogr. Assoc. Meet., Evanston, Illinois, USA, Abstract B1.

Johnson, C. K. (1977). Proc. Am. Crystallogr. Assoc. Meet., Asilomar, CA, USA, Abstract JQ6.

Karle, I., Karle, J., Mastropaolo, D., Camerman, A. & Camerman, N. (1983). *Acta Cryst.* **B39**, 625-637.

Karle, J. & Hauptman, H. (1958). *Acta Cryst.* **11**, 264-269.

Konstandinos, K. (1992). MSc Thesis, Queen's Univ., Kingston, Canada. In preparation.

Langs, D. A. (1988). *Science*, **241**, 188-191.

Langs, D. A. & DeTitta, G. T. (1975). *Acta Cryst.* **A31**, S16.

Lathrop, R. H., Webster, T. A. & Smith, T. F. (1987). *Commun. ACM*, **30**, 909-921.

Leherte, L., Fortier, S. & Glasgow, J. I. (1992). Proc. Am. Crystallogr. Assoc. Meet., Pittsburgh, Pennsylvania, USA, Abstract PA98.

Main, P. (1976). In *Crystallographic Computing Techniques*, edited by F. R. Ahmed. Copenhagen: Munksgaard.

Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E. & Cedergren, R. (1991). *Science*, **253**, 1255-1260.

Martin, T. P., Hung, H.-K. & Walmsley, C. (1992). Proc. 3rd Int. Conf. Database Expert Syst. Appl., Valencia, Spain.

Minsky, M. (1975). *The Psychology of Computer Vision*, edited by P. H. Winston, pp. 211-277. New York: McGraw-Hill.

Peschar, R. & Schenk, H. (1987). *Acta Cryst.* **A43**, 513-522.

Peschar, R. & Schenk, H. (1991). *Acta Cryst.* **A47**, 428-440.

Qian, N. & Sejnowski, T. S. (1988). *J. Mol. Biol.* **202**, 865-884.

Riesbek, C. K. & Schank, R. C. (1989). *Inside Case-Based Reasoning.* Hillsdale: Lawrence Erlbaum Associates.

Rooman, M. J. & Wodak, S. J. (1988). *Nature (London)*, **335**, 45-49.

Sheldrick, G. (1990). *Acta Cryst.* **A46**, 467-473.

Smith, V. H. Jr, Price, P. F. & Absar, I. (1977). *Isr. J. Chem.* **16**, 187-197.

Taylor, R. & Allen, F. H. (1992). In *Structure Correlation*, edited by H.-B. Bürgi & J. D. Dunitz. Weinheim: VCH Publishers.

Thompson, K. & Langley, P. (1991) In *Concept Formation: Knowledge and Experience in Unsupervised Learning*, edited by D. H. Fisher & M. Pazzani. Los Altos: Morgan Kaufmann.