# Case-based Reasoning in Molecular Scene Analysis

J.I. Glasgow, D. Conklin and S. Fortier
Departments of Computing and Information Science and Chemistry
Queen's University, Kingston

## 1 Introduction

In recent years, crystallographic studies have generated an explosion of information on molecular structure. The three-dimensional structural information obtained from such studies provides precise and detailed depictions of molecular scenes, an essential starting point for unraveling the complex rules of structural organization and molecular interactions in biological systems. This information is normally stored in databases in the form of unstructured atomic coordinate data. The research described in this paper is concerned with the organization of this structural knowledge into a case base of molecular scenes. The creation of such a case base provides a foundation for a computational strategy for the analysis of new scenes based on the experience embodied in the stored cases.

Molecular scene analysis [FCG+93; GFA92] is concerned with the automated reconstruction and interpretation of crystal and molecular structures. A key feature of this research is the integration of the available knowledge on molecular structures into image reconstruction tools through the techniques of case-based reasoning and machine discovery. Discovered spatial and visual concepts index cases that can be used to anticipate the identity and relative locations of parts in a molecular scene.

This paper describes how existing structural data can be organized so as to permit efficient and rapid retrieval from a case base of molecular scenes. Section 2 overviews the process of molecular scene analysis. In Section 3, issues involved in the representation, indexing, matching, adaptation and evaluation of molecular scenes are presented. The paper concludes with a discussion of related research in this area.

## 2 Molecular Scene Analysis

Marr [Mar82] defined computational vision as the "process of discovering what is present in the world, and where it is". A similar discovery process takes

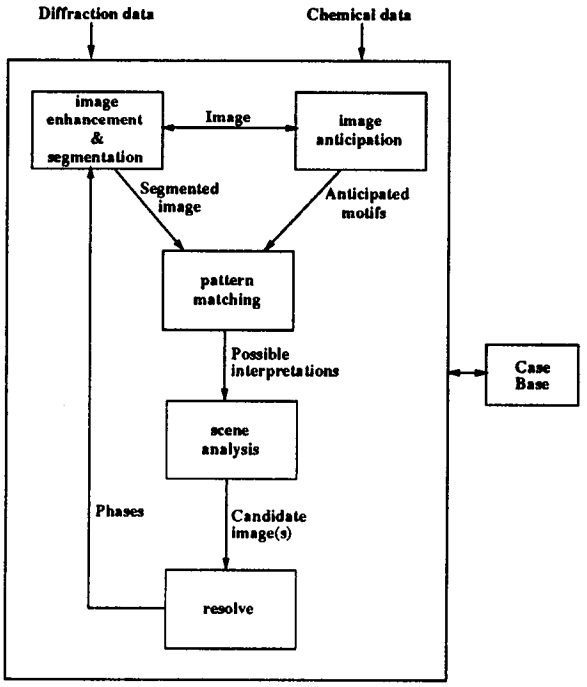


Figure 1: Processes for Molecular Scene Analysis

place in molecular scene analyses. For example, crystal structure determination, molecular structure classification and prediction, and studies of molecular recognition processes, can all be viewed as exercises in molecular scene analysis.

Figure 1 illustrates the five independent, but communicating, processes involved in the reconstruction of a three-dimensional molecular scene from experimental diffraction data [FCG+93]. The *image anticipation* process involves the retrieval of motifs from a case base according to available chemical and structural information. In the *image enhancement & segmentation* process, an experimental electron density map (three-dimensional image of a crystal structure) is subjected to standard noise reduction and density modification routines and segmented into distinct blobs/regions that correspond to structural features of the image. The *pattern matching* process involves the comparison of unidentified features derived from the segmentation process with anticipated motifs retrieved from the case base. In the *scene analysis* process, the possible interpretations are then evaluated using global constraint satisfaction techniques [Kon93]. Finally, in the *resolve* process the structural information determined in the scene analysis process is integrated into the direct methods tools [Hau86] so as to refine and enhance the emerging image.

The process of molecular scene analysis can be expressed as a state space search, where the initial state is an uninterpreted image and the goal state is a comprehensive interpreted image of the scene. The state space for this problem is represented as a set of partially interpreted images, depicted by symbolic arrays containing symbols denoting the molecular parts in a scene. These symbols may correspond to identified or currently unidentified parts. The goal of case retrieval in molecular scene analysis is to find a set of previously determined scenes (cases) which either 1) share visual properties (shape and volume) with an unidentified part in the new image, or 2) share structural features with the image (i.e. have identified parts in a similar configuration). Case-based reasoning is applied primarily in the *image anticipation* process (see Figure 1) of molecular scene analysis.[1] Here, chemical, structural and visual information is used in to retrieve cases containing image motifs for potential pattern matching. Thus, it is important to be able to index the case base according to the visual and spatial qualities of molecular images.

Crystal structure determination, an example of a molecular scene analysis exercise, is an iterative process which can be modeled as the resolution of the three-dimensional image of the atomic arrangement within the crystal. Initially, the electron density maps to be analyzed are of low resolution. Thus, at this stage our case-base reasoning is guided primarily by fuzzy shape information and chemical information (e.g., primary structure). Once some partial structure information has been hypothesized, this information can be used in the *resolve* process to improve the overall image. This is possible because of the Fourier transform relationship between the data and the image formed [FN89]. Once a partially interpreted image has been formed (represented by a partially defined symbolic array), the available three-dimensional structural information can also be used to anticipate the identity of unknown parts of a scene. The rationale here is that parts of an image often appear in similar contexts. Thus, a potential interpretation can be hypothesized based on whether it has occurred previously in a similar context (defined by the interpreted parts of the scene).

The process of crystal and molecular structure determination usually relies on an individual crystallographer's ability to recall, compare and adapt three-dimensional structural motifs of previously interpreted scenes. Because of the chemical constraints placed on molecular images, it is rare that reconstructed scenes do not contain any parts which have not occurred previously in a similar spatial configuration. Thus, efficient retrieval of relevant experiences and the ability to represent, index and pattern match molecular images or motifs, based on their three-dimensional spatial structure, can greatly aid in molecular scene analyses.

Rapidly growing crystallographic databases[2] were conceived in the 1960's

[1] However, it may also be beneficial to use case-based reasoning in image evaluation.

[2] The Cambridge Structural Database contains over 100,000 organo-carbon compounds and the Protein Data Bank contains approximately 700 macromolecules; database sizes grow by more than 10% per year.

and 1970's to contain the explosion of information being generated. For many years, access to the three-dimensional atomic coordinate data, the vital experimental component of the databases, was gained through information retrieval and database analysis techniques applied to bibliographic or chemical data. Basic knowledge, in the form of standardized geometric descriptions of molecules or fragments of molecules, could be generated from retrieved coordinates, but comparative analysis of large numbers of instances was primarily a manual operation. In recent years, an expanding repertoire of statistical and numeric techniques have been introduced to survey, organize and view the geometric data, in order to recognize patterns and relationships [ADT91]. This paper is concerned with the further transformation of the ever-growing databases into case bases that will ultimately provide the foundation to a knowledge-based approach to molecular scene analysis.

# 3 Case-based Reasoning

A crucial component in a computational approach to molecular scene analysis is the ability to retrieve structural motifs from a case base and adapt these to classify parts in a partially interpreted image of a novel molecular scene. Because indexing and matching are based on the three-dimensional structure of a molecular scene, there are unique problems that must be addressed when carrying out case-based reasoning in this domain. In the remainder of this section, we consider the issues of case representation, indexing, matching and adaptation for the problem domain of molecular scene analysis.

## 3.1 Representation of Cases

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. Thus, the representation and indexing schemes for case-based reasoning must capture the fundamental visual and spatial properties of the scenes. Currently, crystallographic databases contain geometric descriptions of the atomic coordinates of a molecular scene. Although this information is sufficient and complete, it is not organized for efficient retrieval and pattern matching based on the structural similarities of scenes. Thus, we incorporate a representational formalism for computational imagery [GP92], which was tailored to facilitate the processes of mental imagery - processes similar to those often used by expert crystallographers in the identification of molecular structures. This framework consists of knowledge representations and primitive operations that are used to depict, transform, scan, pattern match and reason with visual and spatial depictions.

To perform case retrieval based on structural similarity of molecular scenes requires a representation that denotes the meaningful parts of a scene and their

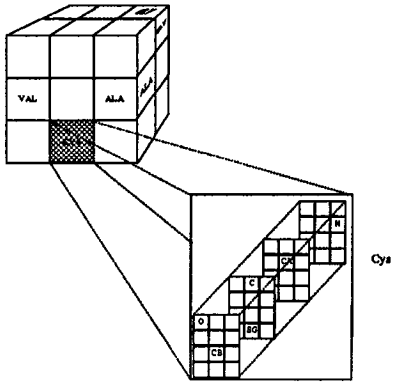


Figure 2: Spatial representation of molecular image

relevant spatial and topological relations. The scheme for computational imagery includes a spatial knowledge representation that provides an explicit symbolic array depiction of an image. As illustrated in Figure 2, the components of a molecular image are denoted as symbolic components of the array [3], while the hierarchical structure of a scene is captured using nested arrays.[4] In molecular scene analysis, such arrays can be reconstructed from the geometric information stored in the existing databases or from the analysis of experimental electron density maps.

Crucial to molecular scene analysis are the location, identification and comparison of molecular features, including molecular shapes at varying levels of resolution. An important step in such analyses is the segmentation of molecules, or molecular fragments, into shape primitives so as to allow for their rapid and accurate recognition. Molecular shape information can be extracted from the three-dimensional electron density maps. The information in these maps, however, is at too detailed a level to allow for efficient pattern matching. Thus, it is essential to transform this visual representation into one that can capture the relevant shape information and discard the unnecessary and distracting details. We are currently investigating and evaluating several models for the segmentation and representation of molecular fragments as shape primitives. These include the critical point mapping method [Joh77], which uses topological properties of the electron density distribution to segment maps and to represent shape, and a generalized cylinder model [Bin71], which represents objects as a hierarchy of segmented parts described as volumetric shapes.

In summary, the contents of a case for a molecular scene includes information

[3] In Figure 2, the symbols in the array correspond to amino acid residues, e.g. alanine (ALA) and cysteine (CYS)

[4] The nested array here denotes the three-dimensional depiction of cysteine containing atomic parts, e.g. nitrogen (N) and carbon (C).

for retrieving visual and spatial knowledge of a molecular scene. This information may be stored explicitly, or it may be stored in a form that allows for the generation of depictions of scenes which can then be inspected to retrieve the relevant information.

## 3.2 Case Retrieval and Adaptation

The ability to perceive spatial similarities and equivalence is fundamental to case-based reasoning in molecular scene analysis. We define two images to be spatially equivalent under a set of relations $R$ if and only if there exists a mapping $f$ between their parts such that if parts are related by $r \in R$ in one image, then $f$ maps them to parts that are similarly related in the other image. Such a mapping can be determined through inspection of the symbolic array depictions.

Spatial equivalence can be used to define the concepts of image similarity for matching of image representations [CG92]. A measurement of spatial similarity between two images is expressed in terms of the minimal number of image transformations (amount of adaptation) needed to bring them into equivalence. Many types of adaptations are possible, such as replacing, deleting or moving a part, or moving parts simultaneously (e.g., rotation). Initial results in studies of molecular similarity suggest that it may be sufficient to consider only part deletion, provided that the relations $R$ are invariant under rotation and translation. Assume that a mapping $f$ is a partial correspondence between two images $s$ and $t$ that preserves a set of relations $R$. Then we can define a similarity valuation for $f$ as the function:

$$S_R(s,t,f) \rightarrow \frac{2\times\#f}{\#s+\#t}$$

where $\#f$ is the cardinality of the domain of $f$, and $\#s$ and $\#t$ are the number of parts in the two images. The similarity valuation $S_R(s,t)$ for two images $s$ and $t$ can then be defined as the maximum value of $S_R(s,t,f)$ over all partial correspondences $f$ that preserve $R$. This similarity measure has the property that structurally equivalent images receive a value of 1; dissimilar images receive a value of near 0.

Central to our method for case retrieval is the abstract *image concept.* This is an image which contains other concepts, rather than instances, as parts. One image *subsumes* another if its parts can be mapped into parts of the other such that all relations are preserved, and subsumption is inductively preserved by the mapping. A case base of images is organized by this subsumption relation. At the leaves of a concept taxonomy are the individual cases; the internal nodes are image concepts. It has been demonstrated that such a model provides an efficient model of structured case retrieval [CG92; Lev92]. It is also semantically well-founded on the classification model of information retrieval [BGN89].

To construct a concept taxonomy, we have developed a structured concept formation system called I-MEM (Image MEMory) [CG92], based on the theory of

description logics [Neb90], and the concepts of image similarity and subsumption described above. This system accepts a stream of cases, incrementally building a taxonomy of image concepts. At any stage, an input case may trigger the formation of a new concept, or may simply be placed under the most specific subsuming concept in the taxonomy. High similarity with an existing case will trigger concept formation. Preliminary results indicate that I-MEM constructs hierarchies that facilitate case retrieval [CG92]. We have previously demonstrated that the I-MEM system for conceptual clustering can also be used to successfully discover spatial concepts in the domain of molecular classification [CFGA92]. Here the classes correspond to meaningful conceptual groupings that can be used to index scenes for retrieval in case-based reasoning.

In conclusion, cases of molecular scenes are retrieved based on the visual and spatial attributes of a scene. Above we have given a brief description of how spatial indices are represented and created. In indexing and matching a case with a partially interpreted scene, it is not only important that the images have similar parts but also that the spatial and topological (bonding) relationships among these parts are preserved. Specialized representations and techniques for spatial analogy and subsumption have been developed to address these issues.

## 3.3 Image Reconstruction and Evaluation

Once the relevant molecular scenes have been retrieved and adapted, they can then be used to predict the identity of a previously unknown part in a molecular scene. In the adaptation step, previously determined scenes are transformed, through functions such as parts deletion or rotation, to come into correspondence with the current scene. Although it is assumed that the scene being analyzed is novel, it is also well known that the structural building blocks of molecules, and particularly of proteins, occur in familiar patterns; a scene can usually be decomposed into meaningful parts that have often been previously experienced in similar configurations and with similar shapes. Thus, a retrieved case can be used to predict the identity of a part based on its visual qualities and/or its context within a neighborhood of interpreted parts. As discussed in Section 2, the prediction of these parts can then be used to resolve the current image which can then be reanalyzed to interpret more of the structure.

Interpretation of parts within a molecular image can be achieved based on the retrieval of a single similar case or on a concept that has been formed using the machine learning techniques described earlier. The retrieval of such experiences can result in one or more of the previously uninterpreted parts of an image being instantiated (i.e., the identity of one or more parts are hypothesized).

Once an image representation has been transformed (identity of parts added) based on scenes retrieved from the case base, image evaluation can be phrased as a constraint satisfaction problem. In particular, an emerging image must adhere to the known crystallographic and chemical constraints imposed by the input data. These include hard constraints, such as chemical constitution, as

well as soft constraints, such as those involved in molecular interactions.

# 4 Concluding Discussion

It is understood that there exists a one-to-one correspondence between the three-dimensional structure of a protein and its amino acid sequence. Considerable research efforts are being devoted to the problem of predicting a protein's structure from its sequence (e.g. [HS91; Kin87; QS88; RW88; ZW89]). Although taking such an approach may be tractable in the future, it is generally accepted that there are currently too few examples to accurately predict structure from the sequence alone [RW88]. We propose an approach to protein structure determination that uses case-based reasoning to anticipate three-dimensional substructures within a molecular scene. Experimental data from crystallographic experiments can then be used to validate the hypothesized parts within the scene. Thus, molecular scene analysis can be phrased as an iterative process using a top-down approach to anticipate molecular parts, working in tandem with a bottom up visual analysis of the image present in the electron density map.

Currently, a great deal of structural information is available, yet not generally exploited, at the outset of a structure determination exercise [ABS87]. Its systematic use is not trivial, in part because the information is available only in the form of coordinate data, defined in terms of an external reference frame. The transformation of the data into cases which can be indexed and matched based on structural similarity is of fundamental importance to the problem of molecular scene analysis.

Similar to PROTOS [Bar89], which diagnoses hearing disorders, our approach to scene analysis applies previous experiences to the task of classification. Given a partially interpreted molecular image, we find potential classifications for structural subimages and use these to predict and evaluate the identity of unknown parts of the scene. The classification problem for molecular scene analysis, however, has some unique features. In particular, it requires the representation and comparison of three-dimensional image representations of molecular structures. To accommodate this, we incorporate specialized representation and machine learning formalisms to store and classify molecular knowledge.

The focus of this paper has been on the use of case-based reasoning in the domain of molecular biology, an area recognized as providing opportunities for integrating a variety of artificial intelligence techniques [Hun92]. The massive amounts of available crystallographic data implies the need for efficient representation, indexing and retrieval techniques applied to chemical and spatial information. Thus, it is of great interest to our project to investigate and evaluate approaches to case-based reasoning and information retrieval.

# References


[ABS87] F.H. Allen, G. Bergerhoff, and R. Sievers, editors. *Crystallographic Databases.* IUCr, Chester, UK, 1987.

[ADT91] F.H. Allen, M.J. Doyle, and R. Taylor. Automated conformational analysis from crystallographic data, 3. Three-dimensional structural database system: implementation and practical examples. *Acta Crystallographica*, B47:50–61, 1991.

[Bar89] E.R. Bareiss. *Exemplar-Based Knoweldge Acquisition: A Unified Approach to Concept Representation, Classification, and Learning.* Academic Press, Boston, MA, 1989.

[BGN89] H. W. Beck, S. K. Gala, and S. B. Navathe. Classification as a query processing technique in the CANDIDE semantic data model. In *Proc. IEEE International Data Engineering Conference*, pages 572–581, 1989.

[Bin71] T.O. Binford. Visual perception by a computer. In *Proceedings of IEEE Conference on Systems and Control*, Miami, Florida, 1971.

[CFGA92] D. Conklin, S. Fortier, J. Glasgow, and F. Allen. Discovery of spatial concepts in crystallographic databases. In *Proceedings of the Machine Discovery Workshop*, pages 111–116, Aberdeen UK, 1992.

[CG92] D. Conklin and J.I. Glasgow. Spatial analogy and subsumption. In Sleeman and Edwards, editors, *Machine Learning: Proceedings of the Ninth International Conference ML(92)*, pages 111–116. Morgan Kaufmann, 1992.

[FCG+93] S. Fortier, I. Castleden, J. Glasgow, D. Conklin, C. Walmsley, L. Leherte, and F. Allen. Molecular scene analysis: The integration of direct methods and artificial intelligence strategies for solving protein crystal structures. *Acta Crystallographica*, D1, 1993.

[FN89] 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.

[GFA92] J.I. Glasgow, S. Fortier, and F.H. Allen. Molecular scene analysis: crystal structure determination through imagery. In L. Hunter, editor, *Artificial Intelligence and Molecular Biology.* AAAI Press, 1992.

[GP92] J.I. Glasgow and D. Papadias. Computational imagery. *Cognitive Science*, 16(3):355–394, 1992.

[Hau86] H. Hauptman. The direct methods of X-ray crystallography. *Science*, 233:178–183, 1986.

[HS91] L. Hunter and D.J. States. Applying Bayesian classification to protein structure. In *Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications*, Miami, Florida, 1991.

[Hun92] L. Hunter. Artificial intelligence and molecular biology. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 866–868, 1992.

[Joh77] C.K. Johnson. Peaks, passes, pales and pits: a tour through the critical points of interest in density maps. In *Proceedings of the American Crystallographic Association Meeting*, Asilomar, California, 1977. Abstract JQ6.

[Kin87] R. D. King. An inductive learning approach to the problem of predicting a protein's secondary structure from its amino acid sequence. In *Proc. EWSL '87*, 1987.

[Kon93] K. Konstantinos. Evaluation of partially interpreted images. Master's thesis, Queen's University, Kingston, Canada, 1993. to appear.

[Lev92] R. Levinson. Pattern associativity and the retrieval of semantic networks. *Computers Math. Applic.*, 23(6):573–600, 1992.

[Mar82] D. Marr. *Vision.* W.H. Freeman and Company: San Francisco, 1982.

[Neb90] B. Nebel. *Reasoning and revision in hybrid representation systems.* Springer-Verlag, 1990.

[QS88] N. Qian and T.S. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Molecular Biology*, 202:865–884, 1988.

[RW88] M.J. Rooman and S.J. Wodak. Identification of predictive sequence motifs limited by protein structure data base size. *Nature*, 335:45 – 49, 1988.

[ZW89] X. Zhang and D. Waltz. Protein structure prediction using memory based reasoning: A case study of data exploration. In *Proceedings of a Workshop on Case-Based Reasoning*, pages 351–355, San Mateo, California, 1989. Morgan Kaufmann Publishers, Inc.