From: ISMB-94 Proceedings. Copyright © 1994, AAAI (www.aaai.org). All rights reserved.

VQLM: A Visual Query Language for Macromolecular Structural Databases

Dawn Cohen[†], Kumar Vadaparty[‡], Bill Dickinson[‡] and Hemanth Salem[‡]

[†]Keck Center for Computational Biology
University of Pittsburgh
Pittsburgh, PA 15260
dcohen@cs.pitt.edu

[‡]Department of Computer Engineering and Science
Case Western Reserve University
Cleveland, OH 44106
kumarv, billd, salem@alpha.ces.cwru.edu

Abstract

Databases of macromolecular structures allow researchers to identify general principles of molecular behavior. They do this by providing a variety of data obtained under a number of different experimental conditions. Many new tools have been developed recently to aid in exploratory analysis of structural data. However, some queries of interest still require considerable manual filtering of data. In particular, studies attempting to make generalizations about complex arrangements of atoms or building blocks in macromolecular structures cannot be approached directly with existing tools. Such studies are frequently carried out on only a few structures or else require a labor-intensive process. To address this problem, we have developed a visual language, VQLM (Visual Query Language for Macromolecules). A query is formulated in this language by drawing an abstract picture of substructures to be searched for in the database and specifying constraints on the objects in them. To illustrate the usefulness of our language, we show how to encode a number of queries that were found scientifically interesting in the published literature in molecular biology. VQLM relies on VQL, a new database language, as its underlying engine for database retrieval and computation. We believe that VQLM will make macromolecular structural data more accessible to scientists, enabling faster and deeper data analysis.

Keywords: macromolecular structure, object-oriented databases, graphical user interfaces

1. Introduction

Macromolecular structural databases which contain large numbers of crystal structures have made it possible to study the general principles underlying the behavior of biological molecules. Until several years ago, the macroolecular structure databases mainly provided flat atomic coordinate files, from which a researcher would compute any higher level structural information necessary for a particular study (e.g. PDB (Bernstein *et al.* 1977)). However, some databases provided higher level information, stored in relational and object-oriented databases (e.g. NDB (Berman *et al.* 1992), CSD (Allen *et al.* 1979), P/FDM (Gray *et al.* 1990)). Derived values such as bond lengths and angles, torsion angles, base morphology parameters and virtual bonds are stored in some of these. An object-oriented class library (Chang *et al.* 1994) has been developed for manipulating and computing with this higher level information. In addition, it has become possible to access subsets of data selectively, considering, for example, only data for residues of a single type. However, even with the greatly improved representation of information, some queries of biological interest remain elusive and to answer them still requires considerable manual, labor-intensive and repetitive processing of available data. In particular, it is almost impossible to search for complex arrangements of objects each of which has a specific set of properties.

In this paper we propose a language called VQLM (for Visual Query Language for Macromolecules) which offers the possibility of greatly facilitating future studies in macromolecular structure. In this language, a "picture" of the data to be selected from the database is drawn by the researcher. The components of the picture correspond to objects of the domain and domain-specific relationships between them. Each picture corresponds to a kind of prototypical example of an arrangement of objects to search for in the database. Properties of the individual objects in the picture may be constrained, in order to obtain a more specific set of instances of the given arrangement of objects. The language resembles the QUEST3D (Allen 1992) portion of the CSD. However, VQLM provides tools (objects and constructs) specifically for facilitating queries of *macromolecular* structure (rather than for small molecule fragments).

The power of this language arises from two main sources. First of all, VQLM allows the researcher to refer to object relationships that have some meaning in the domain, such as a pairing between two bases. This is in contrast with most query languages which require that data be accessed in terms of its syntactic form, for example by a join between two relations or set membership. The second source of power lies in the *visual* nature of its queries. The researcher may

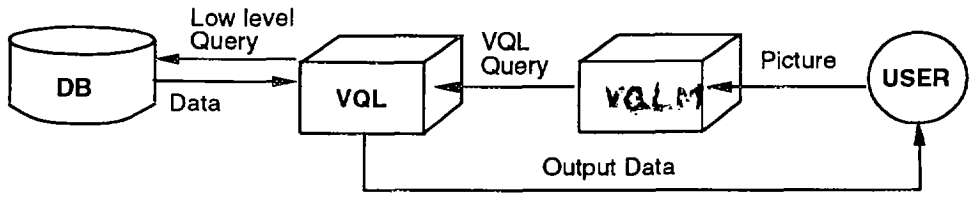

Figure 1: Accessing data through VQLM

state a query as a picture that corresponds directly to what is intended. Again, this is in contrast with the often confusing process of formulating a query as a flat string of logical conditions. Clearly the queries must be compiled into such a flat logical expression, but this work can be managed by a query translator, rather than requiring the researcher to conform to the needs of the database system.

VQLM relies on VQL (Visual Query Language) for its underlying query processing. The relationship between the two is shown in Figure 1. VQL is a declarative query language that incorporates aspects of Prolog (or more accurately, Datalog) and QBE. It extends these languages to behave as a front-end for object oriented databases. An important characteristic of VQL is that it uses a special construct called *restricted universal quantifier*. It has been shown in (Vadaparty, Aslondogan, & Ozsoyoglu 1993) that this construct is quite powerful: it can simulate easily a number of useful constructs such as grouping, MIN, MAX, etc. Thus, this construct gives VQL versatility. It has also been shown that VQL is strictly more powerful than Datalog, even with stratified negation.

An advantage of VQL is that it allows one to "program pictorially", which reduces the chances of error. Furthermore, VQL translates its queries into a language that can be interfaced with object-oriented or relational database management systems. Thus, the user does not need to know details of the database management system that is being used. This provides transparency.

VQLM is under development. This paper is primarily meant to present the language and show how it could be used in structure studies, rather than to report findings using it. The queries illustrated and discussed here have *not* been run. We currently focus on queries about nucleic acid structure. Many of the primitives are specific to nucleic acid structure. Others are applicable to nucleic acid, protein or general molecular structure. In the future, primitives specifically for describing protein structure will be formalized.

The rest of this paper is organized as follows. Section 2 motivates this work with some queries that have previously been addressed in the molecular biology literature and shows how they would be encoded in VQLM. Section 3 describes the language in terms of the objects and operators that can be used to build queries and the interface for making use of these primitives. Section 4 concludes with a discussion of the contributions of this work and some directions for future work.

2. Example Queries

In this section we briefly describe several studies from the molecular biology literature and show how they could have been encoded in VQLM. Each query is explained in detail, though the primitives used for constructing the queries are described later, in Section 3.

Constructing a Query – B-DNA Guanine Hydration

In Figure 2 we illustrate a simple query representing "Find the coordinates of waters within hydrogen bonding distance of a guanine base in a B-DNA molecule." This query and others very similar to it were integral to several studies including (Berman *et al.* 1988), (Schneider *et al.* 1993) and (Cohen, Kulikowski, & Berman 1993). Though not technically challenging to compute, this query presented some difficulty due to a lack of appropriate tools for computing it.

The VQLM version of this query is shown in Figure 2 and is explained as follows. A guanine base named B1 is connected to a water named W1 by a distance relationship drawn as a line. This specifies that we are interested in waters within some distance *d* of guanines. Constraints on the distance relationship are abbreviated by a variable *d* in the figure. In the *Constraints* box, d is set to the generally accepted hydrogen bonding range of 2.6-3.2$\AA$. The molecule M1 containing the base is drawn around it and in the *Constraints* box, its DNA-type is shown to be "B". (Please note that in this paper, an attribute of an object will be referred to with the notation **object-name.attribute-name.**) A crystal is drawn around the molecule and the water to denote that both must come from the same crystal. The water is highlighted in the figure to denote that it has just been selected. The *Symmetry* menu has *Include Symmetry Related* highlighted to denote that waters (i.e. the selected W1) outside the asymmetric unit should be included in the query computation. The output table is named as *WG-Tbl* and includes three columns of data, corresponding to the x, y and z coordinates of the selected waters. The output data columns are headed by the titles X, Y and Z respectively.

In order to construct this query, the following steps

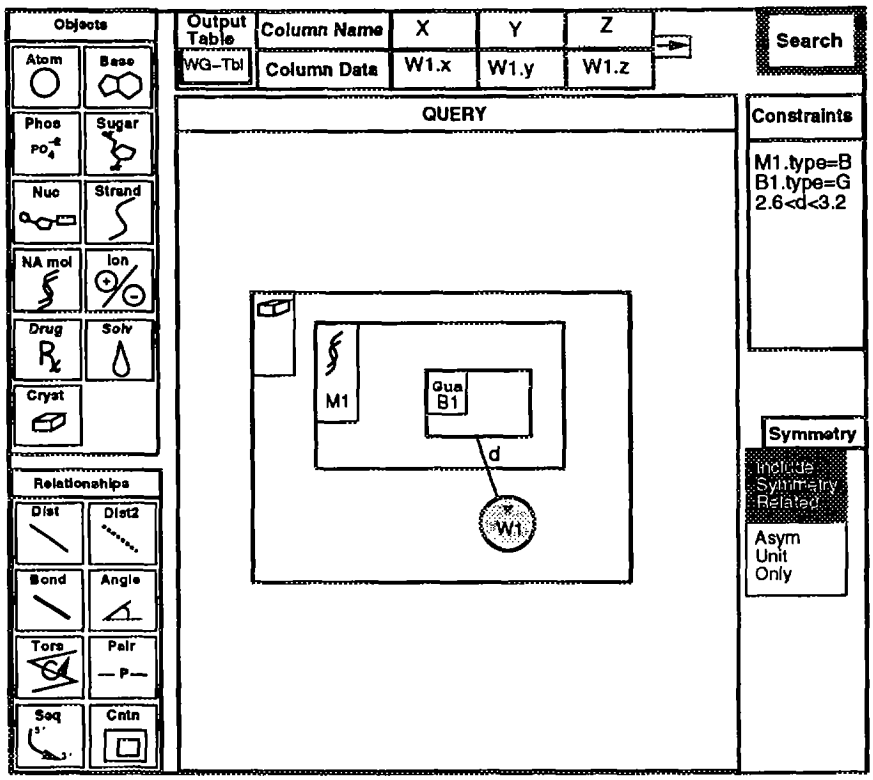

Figure 2: The VQLM interface with guanine hydration query

might have been taken (though there are other possibilities). First, by clicking on the *Base* icon, a base may be produced in the *Query* box. It is named B1 by the user. By clicking on the base, a menu of its attributes will be provided which allows constraints on any of its attribute values to be specified. In this query, the *base-type* attribute is set to "G" which is reflected in the *Constraints* box and on the base itself (its labeling turns to "Gua"). Next, by clicking on the *Solvent* icon, a solvent molecule is produced and by specifying its *solvent-type* attribute with the value "water", it is constrained to be a water. Next a distance relationship between the water and the base may be specified by clicking on the *Distance* icon and then clicking on the two objects. This will cause a prompt for a distance range to appear on the screen which the user must fill in. Next, by clicking on the *Nucleic Acid Molecule* icon, a molecule appears in the *Query* box. By sizing it appropriately and dragging it so that it surrounds B1, the user asserts that the guanine is contained in the molecule. Alternatively, by clicking on the *Containment* icon and then clicking on the molecule and the base, it is asserted that the base is a subpart of the molecule, and it appears inside the molecule object. Then, by clicking on the molecule and obtaining its attribute menu, the DNA-type of the molecule can be set to "B". Similarly, by clicking on the *Crystal* icon, a crystal object appears in the *Query* box and by using the *Containment* icon, the user asserts that both the DNA molecule and the water are subparts of the same crystal. Finally, by clicking on the water to select it, and then selecting *"Include Symmetry Related"* from the *Symmetry* menu, it is specified that symmetry generated waters should be included in the query computation. Clicking on the *Search* icon causes the output table to be computed.

The desired output is specified as follows. Initially, there are no output columns. By clicking on the arrow, a new, empty column is produced. The user then specifies a header for the column in the *Column Name* box and an attribute of one of the objects in the *Column Data* box. Thus, to obtain the desired water coordinates, the column headers X, Y and Z are specified, and the x, y and z coordinates of the water object are used as the output data.

Water Bridges in DNA Crystals

There is a fairly extensive literature documenting crystallographers' interest in hydrogen-bonding networks of waters around nucleic acids. The simplest network, the *water bridge* has received considerable attention. This arrangement consists of two DNA atoms hydrogen-bonded to a single water molecule. There has been some discussion in the biophysical community of conditions under which particular kinds of bridges form (e.g. base-water-base bridges, phosphate-water-

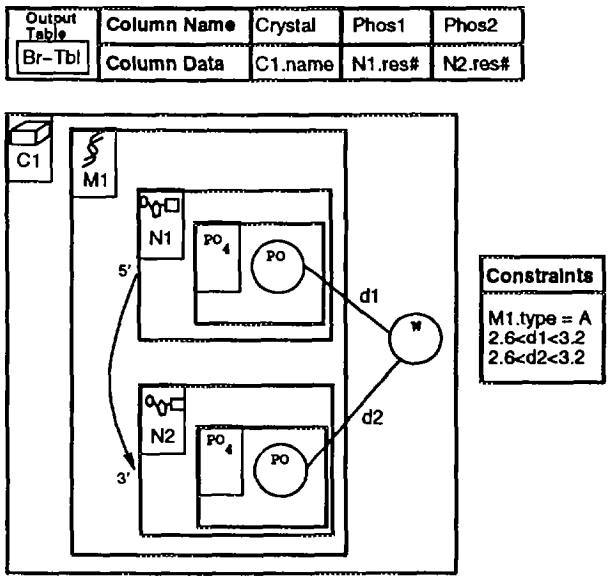

Figure 3: Query for phosphate-water-phosphate bridges in A-DNA

phosphate bridges, etc.) It is believed that they contribute to the stability of crystal structures and they have been cited in explanations of several phenomena of DNA behavior. See (Westhof 1993) for a survey of the literature on DNA-water bridges. In VQLM, it is quite simple to write a query for finding various types of water bridges, in contrast to the current laborious procedure for finding them.

Figure 3 shows a query representing "Find the crystals and residue numbers of all pairs of phosphates in adjacent nucleotides in A-DNA molecules that have a water hydrogen-bonded to oxygens in both phosphates". It has been claimed (Saenger, Hunter, & Kennard 1986) that A-DNA tends to form in a low humidity environment because it allows the relatively scarce water molecules to be shared between neighboring phosphates. In contrast, the large distance between phosphates in B-DNA makes bridging impossible so that its phosphate oxygens must be surrounded with more waters. This provides an explanation of why B-DNA has mainly been observed in crystals grown at relatively high humidity. This argument was based on an examination of several crystal structures by the authors of (Saenger, Hunter, & Kennard 1986). It could be further tested by looking at the relative numbers of A-DNA and B-DNA phosphate–water–phosphate bridges in the entire database, using the query in Figure 3.

Molecular Packing

How molecules interact with or contact each other in crystals is a topic of considerable interest to crystallographers. Again, each crystal structure tends to be analyzed by itself in detail and general patterns occurring in many structures can only be found either by the fact that several authors notice and report the same phenomenon or with a difficult analysis. One somewhat more general study looked at packing in a variety of Z-DNA molecules (Schneider *et al.* 1992). This study mentions that in many of these structures, the terminal O3' of only one strand is very close to a phosphate oxygen of a specific nucleotide in a symmetry related molecule. This fact could have been discovered easily with the VQLM query shown in Figure 4. This query corresponds to "Find close pairs of atoms in symmetry related Z-DNA molecules and report their atom types and the crystal and residue they are contained in".

Sequence-Structure Interactions

It is common knowledge that sequence and function are integrally related for macromolecules. The exact mechanism for this is not known, however. Thus, as an intermediate step, some crystallographers study the relationship between local sequence and local structure. (Drew, McCall, & Calladine 1990) cites many studies on this topic. An example of a query from one such study (Nelson *et al.* 1987) is shown in Figure 5.

The query in Figure 5 represents "Find the crystals containing an adenine that is between two other adenines and show the crystal name, residue number of the middle adenine and the propellor twist of its base-pair". It was claimed in (Nelson *et al.* 1987) that an adenine-thymine pair surrounded by adenine-thymine pairs on both sides exhibits higher values of propellor twist than pairs in other sequences do. This is explained by the fact that the high propellor twist allows a stabilizing hydrogen bond to form between an adenine and a diagonally opposite thymine. This ar-

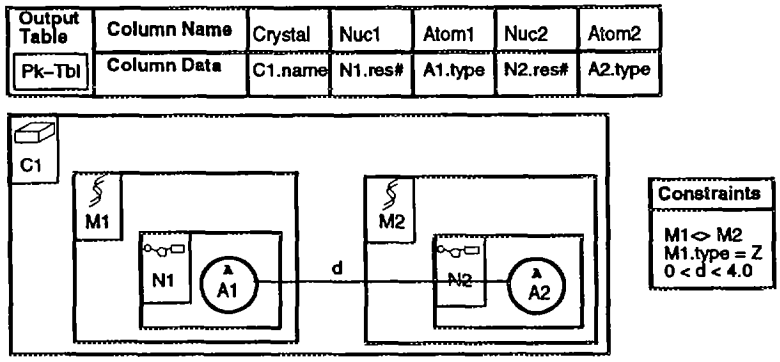

Figure 4: VQLM query examining packing of Z-DNA molecules

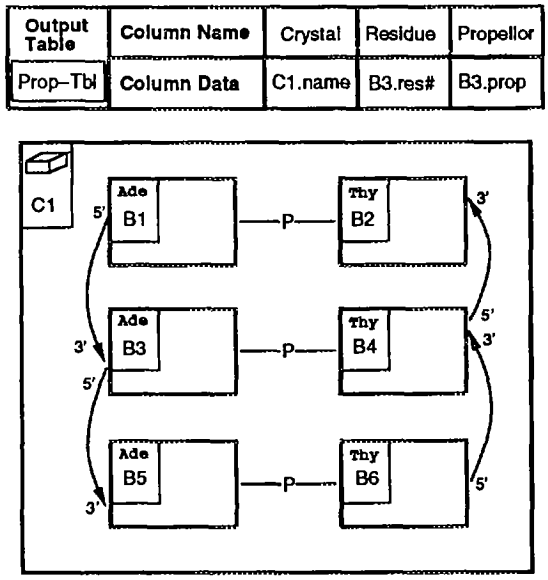

Figure 5: Query for propellor twist of A's in runs of A's

gument from (Nelson *et al.* 1987) was based on an examination of *one* crystal structure containing the sequence CGCAAAAAACGC. In order to test this further, it might be desirable to run the query of Figure 5 on the entire database and compare the distribution of obtained propellor twists with those obtained for other sequences.

3. A Visual Query Language for Structure

The Interface

In Figure 2 the VQLM interface to be implemented is illustrated.

In VQLM, a query specifies a set of objects, a set of constraints on them and between them, and a set of values to be output. Each object has a set of attributes, any of whose values may be constrained in the query. In addition, the query may constrain an object to be in a set of relationships with other objects. Output is specified as a set of attributes of the objects in the query.

VQLM is meant to provide a graphics program "look-and-feel" for creating queries. Clicking on an object icon will create an object of the specific type in the *Query* box. Clicking on a relationship icon will allow the user to click on objects of the appropriate types to specify that these objects are in the given relationship. Once the objects in a relationship have been specified, the connection between them will be drawn in the *Query* box, the specified constraint will appear in the *Constraints* box and it will be translated into a new portion of the VQL query. A VQLM query can be thought of as a conjunction of the relationships drawn between objects and the constraints on attribute values of the objects. (It will be possible to specify disjunction within the attribute value constraints.)

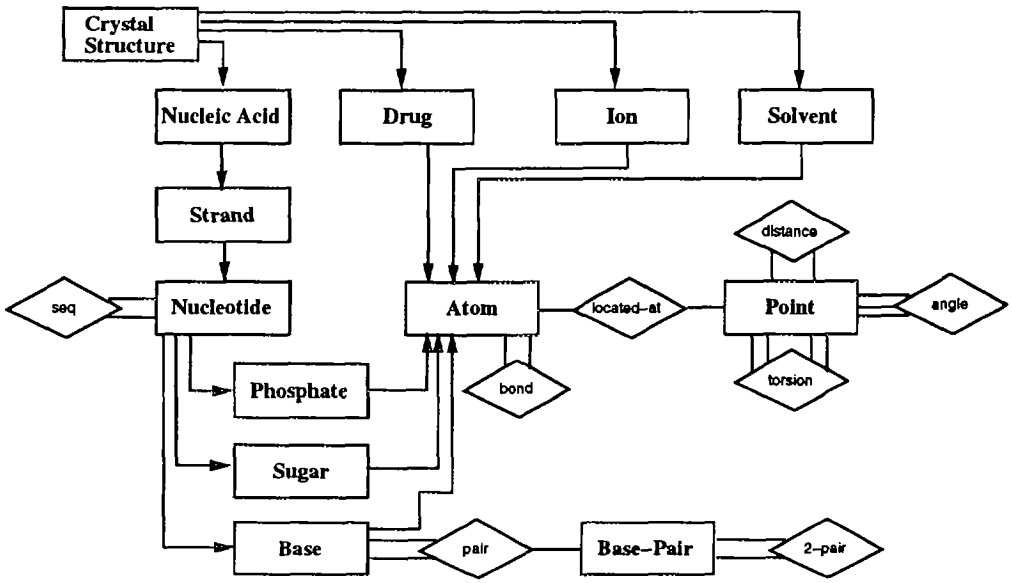

Figure 6: Entity-Relationship Diagram for VQLM

Database organization

The database organization for VQLM is modeled closely on the natural structure of nucleic acid crystals. The entity-relationship (ER) diagram for the database is shown in Figure 6. In that diagram, relationships not otherwise labeled represent the "contains" relationship. Lines represent one-one relationships, while arrows represent one-many relationships.

The primary feature of the ER diagram is the hierarchical structure describing a nucleic acid crystal. The main components of the crystal structure are nucleic acid molecules, drug molecules, ions and solvents. A nucleic acid molecule is composed of a set of strands, which are in turn composed of a set of nucleotides. Nucleotides are related to other nucleotides via the molecular sequence ("seq" relation). Nucleotides are composed of a phosphate group, a sugar, and a base. Each of these, in turn, are composed of atoms. Bases are related to other bases via pairing. Base pairs are related to other base pairs via "double pairing" (the double pairs are necessary for the purposes of describing some base-morphology parameters such as roll, tilt and twist). The non-nucleic acid molecules do not have common substructures, and are composed only of atoms.

Atoms are located at a point in the crystal structure coordinate system (its x, y, z coordinates). A point in the crystal is considered a separate entity. This distinction between points and atoms will eventually make it possible to express more general geometric relationships between arbitrarily defined locations. For example, it may be desirable at some point to implement a method for computing the centroid of an object. Clearly, the centroid does not correspond to any single atom, and must be represented by a point. This would make it possible, for example, to compute distances between centroids of neighboring bases, as a measure of their stacking.

Each type of object has a set of attributes. These attributes are somewhat akin to the columns of data in a relational database table. Some of these represent the raw data of crystallographic analysis, such as the type and coordinates of an atom or unit cell dimensions of the crystal. Others represent derived information such as bond angles, backbone torsion angles, virtual bonds and angles and base-morphology parameters like propellor-twist.

Relationships Between Objects for Building Queries

Substructures to be searched for in the database are described in VQLM by drawing objects and specifying relationships between them. In this section, we describe the main relationships that can be specified and their meanings. They are illustrated in Figure 7.

The relationships between objects fall into two main categories: the geometric and the structure-oriented. The geometric relationships are *distances* between pairs of points, *angles* between triples of points and *torsion angles* between sets of four points. The structure-oriented relationships correspond to domain-specific elements of structure, including *bonding* between atoms, base or nucleotide *pairing*, base or nucleotide *sequence*, and *containment* (or part-subpart relationship).

The structure-oriented relationships are easily represented as attributes of the objects in the relationship. For example, drawing two bases connected a base pair relationship might be translated into an fragment of a

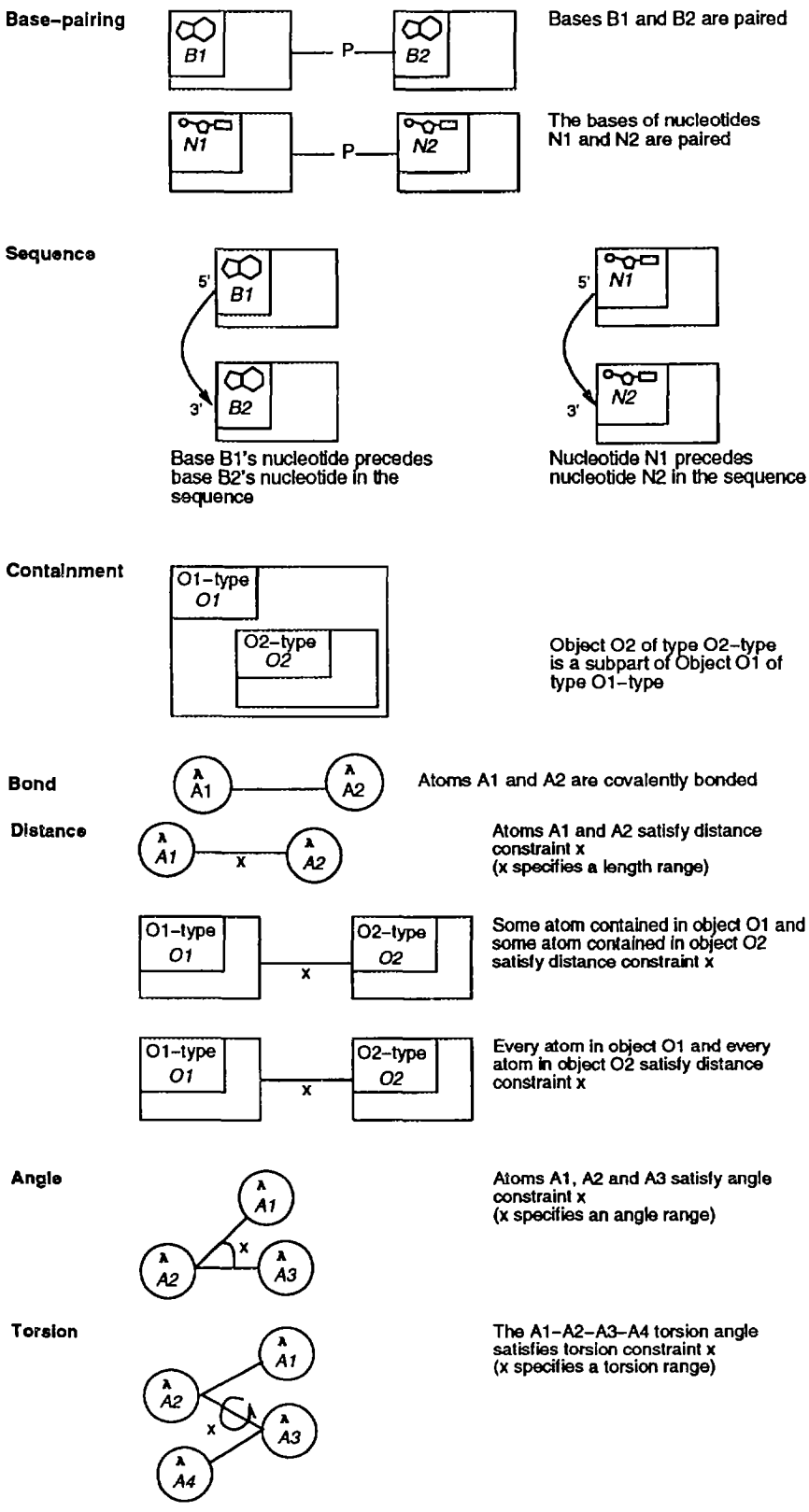

Figure 7: VQLM's relationships between objects

query like "Select b1 in *bases* and b2 in *bases* where *b1.paired-base* = *b2*".

The geometric relationships are more easily computed as functions or methods. For example, drawing two bases connected by a distance (Dist) relationship with distance constraint d = [0-4Å] might be translated into a query fragment like "Select b1 in *bases* and b2 in *bases* where for some a1 in *b1.atomlist* and for some a2 in *b2.atomlist* distance (a1, a2) is between 0 and 4 Å". The Dist2 relationship is a bit more complicated, and relies on VQL's restricted universal quantifier. A picture of two bases connected by the Dist2 relationship with constraint d would be translated roughly as follows "Select b1 in *bases* and b2 in *bases* where for all a1 in *b1.atomlist* and for all a2 in *b2.atomlist* distance(a1, a2) is in the range d".

The geometric relationships provide some computational challenges. The naive approach to distance queries would be to compute all pairs of distances. This approach is very time consuming. It requires $O(n^2)$ comparisons, where n may be the number of atoms in the unit cell of a crystal (for packing queries involving symmetry related molecules). The field of computational geometry ((Preparata & Shamos 1985), (Aurenhammer 1991)) provides the basis for more efficient approaches. Distance queries in the plane have been extensively studied. Algorithms have been given by Bentley and Maurer (Bentley & Maurer 1979), Chazelle et. al. (Chazelle *et al.* 1986), and Aggarwal et. al. (Aggarwal, Hansen, & Leighton 1990). The best approaches run in $O(nlogn)$ time. The type of 3D distance queries required here have not been as extensively studied. Angle queries are between three points and torsion angles are between four points. The naive approach to these queries would be to compute angles and torsions between all possible sets of atoms. This approach requires even more time than the naive distance query algorithm. It is an $O(n^3)$ algorithm for angles and $O(n^4)$ algorithm for torsion angles. Little work in computational geometry has been done with angle queries. Thus, part of our research is aimed at devising efficient 3D distance, angle and torsion query algorithms.

4. Conclusion

VQLM provides a simple graphical interface for making ad-hoc queries on molecular structure databases. Several examples drawn from the biophysical literature illustrate the ease with which fairly complex queries may be specified. Queries may be stated concisely in this language because primitives that are meaningful in the domain are provided. In addition, the pictorial representation may serve as a sort of mnemonic for the expert, allowing them to express a query as they visualize it.

Currently, VQLM is under development. Immediate plans include a full implementation of the functionalities described in this paper. In addition, there are several other primitives which have not been mentioned in this paper that will be made available to the user. Implementation is planned in C++ and versions are expected to be developed both for a 486-based PC under Windows and for X-windows.

Longer range plans include development and implementation of a set of primitives and objects appropriate for querying protein structures. New objects will clearly include amino acids (with side-chain and main-chain subparts), helices, chains of beta-sheet, turns, random coil, and prosthetic groups. New relationships will include phi and psi angles, parallel and anti-parallel hydrogen bonding between beta-chains, disulfide linkages and other types of cross-linking.

Acknowledgments

DMC became aware of many of the issues discussed in this paper during her doctoral research under Prof. Helen Berman. Many thanks to her for this and for sharing her knowledge of DNA structure analysis and her experience in developing molecular structure databases with all of us. Thanks to John Westbrook for several pointers on molecular structure databases. Also thanks to Bohdan Schneider for information on DNA crystallography and helpful suggestions for the design of VQLM and comments on earlier drafts of the paper. Thanks to Bruce Buchanan for support and encouragement of the project. This work was funded in part by a postdoctoral fellowship for DMC from the W.M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University and the Pittsburgh Supercomputing Center. KV and HS were supported under grant IRI93-09320 from the National Science Foundation.

References

Aggarwal, A.; Hansen, M.; and Leighton, T. 1990. Solving query retrieval problems by compacting voronoi diagrams. In *Proceedings of the 22nd Annual ACM Symposium on STOC*, 331–340.

Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; and Watson, D. G. 1979. The Cambridge crystallographic data centre: Computer-based search, retrieval, analysis and display of information. *Acta Cryst.* B35:2331–2339.

Allen, e. a. 1992. *CSD System Documentation.* Cambridge Crystallographic Data Centre, Cambridge, U.K.

Aurenhammer, F. 1991. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys* 23(3):345–405.

Bentley, L. J., and Maurer, H. A. 1979. A note on euclidean near neighbor searching in the plane. *Inf. Process. Lett.* 8:133–136.

Berman, H. M.; Sowri, S.; Ginell, S.; and Beveridge, D. 1988. A systematic study of patterns of hydration

in nucleic acids: (I) guanine and cytosine. *Journal of Biomolecular Structure and Dynamics* 5:1101–1110.

Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.-H.; Srinivasan, A. R.; and Schneider, B. 1992. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal* 69:751–759.

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Mayer, E. F. J.; Bryce, M. D.; Rodgers, J. R.; Kennard, O.; Simanouchi, T.; and Tasumi, M. 1977. The protein data bank. *Journal of Molecular Biology* 112:535–542.

Chang, W.; Shindyalov, I. N.; Pu, C.; and Bourne, P. E. 1994. Design and application of pdblib, a C++ macromolecular class library. *CABIOS.*

Chazelle, B.; Cole, R.; Preparata, F. P.; and Yap, C. 1986. New upper bounds for neighbor searching. *Information and Control* 68:105–124.

Cohen, D. M.; Kulikowski, C.; and Berman, H. 1993. Knowledge-based generation of machine learning experiments: Learning with DNA crystallography data. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 92–100.

Drew, H. R.; McCall, M. J.; and Calladine, C. R. 1990. New approaches to DNA in the crystal and in solution. In *DNA Topology and its Biological Effects.* Cold Spring Harbor Laboratory Press. 1–56.

Gray, P. M. D.; Paton, N. W.; Kemp, G. J. L.; and Fothergill, J. E. 1990. An object-oriented database for protein structure analysis. *Protein Engineering* 3.

Nelson, H. C. M.; Finch, J. T.; Bonaventura, F. L.; and Klug, A. 1987. The structure of an oligo(dA)oligo(dT) tract and its biological implications. *Nature* 330(6145):221–226.

Preparata, F. P., and Shamos, M. I. 1985. *Computational Geometry: An Introduction.* New York: Springer-Verlag.

Saenger, W.; Hunter, W. H.; and Kennard, O. 1986. DNA conformation is determined by economics in the hydration of phosphate groups. *Nature* 324:385–388.

Schneider, B.; Ginell, S. L.; Jones, R.; Gaffney, B.; and Berman, H. M. 1992. Crystal and molecular structure of a DNA fragment containing a 2-aminoadenine modification: The relationship between conformation, packing and hydration in Z-DNA hexamers. *Biochemistry* 31:9622–9628.

Schneider, B.; Cohen, D. M.; Schleifer, L.; Srinivasan, A. R.; Olson, W. K.; and Berman, H. M. 1993. A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *The Biophysical Journal.*

Vadaparty, K.; Aslondogan, Y.; and Ozsoyoglu, G. 1993. Towards a synthesis in visual database access. In *Proceedings of the 1993 ACM SIGMOD Conference.*

Westhof, E. 1993. Structural water bridges in nucleic acids. In Westhof, E., ed., *Water and Biological Macromolecules.* Boca Raton: CRC Press. 226–243.