Journal of Structural Biology **125,** 112–122 (1999)
Article ID jsbi.1999.4102, available online at http://www.idealibrary.com on IDEAL®

A Framework for Querying a Database for Structural Information on 3D Images of Macromolecules: A Web-Based Query-by-Content Prototype on the BioImage Macromolecular Server

P. A. de Alarcón,* A. Gupta,† and J. M. Carazo*

**Centro Nacional de Biotecnología-CSIC, Campus Universidad Autonoma, Cantoblanco, 28049 Madrid, Spain; and †Center for Applied Computational Science and Engineering, San Diego Super Computer Center, University of California at San Diego, La Jolla, California 92093-0505*

Received December 4, 1998, and in revised form February 8, 1999

Nowadays we are experiencing a remarkable growth in the number of databases that have become accessible over the Web. However, in a certain number of cases, for example, in the case of BioImage, this information is not of a textual nature, thus posing new challenges in the design of tools to handle these data. In this work, we concentrate on the development of new mechanisms aimed at "querying" these databases of complex data sets by their intrinsic content, rather than by their textual annotations only. We concentrate our efforts on a subset of BioImage containing 3D images (volumes) of biological macromolecules, implementing a first prototype of a "query-by-content" system. In the context of databases of complex data types the term query-by-content makes reference to those data modeling techniques in which user-defined functions aim at "understanding" (to some extent) the informational content of the data sets. In these systems the matching criteria introduced by the user are related to intrinsic features concerning the 3D images themselves, hence, complementing traditional queries by textual key words only. Efficient computational algorithms are required in order to "extract" structural information of the 3D images prior to storing them in the database. Also, easy-to-use interfaces should be implemented in order to obtain feedback from the expert. Our query-by-content prototype is used to construct a concrete query, making use of basic structural features, which are then evaluated over a set of three-dimensional images of biological macromolecules. This experimental implementation can be accessed via the Web at the BioImage server in Madrid, at http://www.bioimage.org/qbc/index.html.

Key Words: **query-by-content; image databases; structural biology; electron microscopy; BioImage.**

1. INTRODUCTION

Access to computational systems and biological databases currently available on the Internet, and in particular over the WWW, is becoming an important tool for biologists as well as biochemists. Traditional examples are the well-known databases of sequences of nucleic acids (GenBank and the EMBL Data Library) and those for protein sequences (SWISSPROT and PIR), although there are many more available (for a review, see the *Nucleic Acids Research* Special Issue on Databases, January 1999).

The term "query-by-content" has seldom been used in the context of biological databases, making a contrast with its generalized use in other fields, such as generic databases of complex types. However, some of the functionality implied by this term has been in common usage in biology. As a simple example let us consider the quite normal situation of obtaining a new gene sequence and then accessing GenBank to query for all the other sequences present in the database that are somehow similar to the query sequence. Certainly, the query will not be made over textual attributes of the sequence description, but most likely algorithms such as FASTA or BLASTA will be run (for a review see Smith, 1994) and, as a result, a ranking of those similar sequences will be obtained. This usage represents, therefore, an example of a form of query-by-content broadly used in biology.

In the context of structural databases, one of the most used is PDB (Protein Data Bank; see the work by Sussman *et al.* (1998)), which stores thousands of atomic-resolution structures of proteins and nucleic acids and whose growth rate is constantly rising. Data sets in PDB contain a textual description followed by a list of coordinates of atoms (more complete descriptions are also possible). In contrast to sequence databases, filling out a form with textual fields, such as key words, organism, resolution, is the normal form of querying in PDB. Thus, it is not strange that different authors make such an effort to develop more sophisticated accessing systems based on the structural information itself stored in PDB.

1047-8477/99 $30.00

Such applications can be considered query-by-content systems since they attempt to search for the structural information of a PDB entry (i.e., the coordinates of the atoms that form a given protein or nucleic acid). Considering that this work will introduce a query-by-content approach on another structural database (BioImage as restricted to 3D images of biological macromolecules), a review of the main methods aimed at querying PDB by accessing the informational content of its three-dimensional structural information will be presented in the following section.

Still within the context of structural databases, the next step in biological complexity is the 3D images of biological macromolecules contained in BioImage. However, a 3D image is quite different in nature from an ordered string of atomic coordinates. As a consequence, developments in the field of query-by-content applied to generic image databases become especially relevant for its use in BioImage. Based on this consideration, a brief overview of query-by-content systems working on image databases will be provided in another section of this work.

The outline of this paper is as follows. After the Introduction two sections focus on presenting an overview of the approaches carried out so far on the topic of query-by-content first in PDB and then in general image databases. After this, an analysis will be performed on some of the specificities associated with the analysis of biological systems. Section 5 will cover a number of important issues associated with the system layout as well as the implementation of this first prototype of query-by-content on 3D images of BioImage. More specific considerations related to the interfaces design will be discussed under Section 6. A concrete query-by-content prototype will then be constructed and evaluated over a version of BioImage, and eventually, the results will be presented in Section 7. Finally, general conclusions as well as perspectives will be the subject of Section 8.

2. OVERVIEW OF QUERY-BY-CONTENT APPROACHES ON PDB

Similarity searching in databases of 3D structures is considered to be a field of great importance since it could help in the discovery of novel biologically active molecules and investigations of the relationship between proteins' structure and their function. In essence, some of the most relevant methods to be reviewed in this section develop approaches to analyze the structure of molecules by using interatomic distance matrices. One of the most challenging issues is to choose an effective measure to quantify the degree of structural resemblance between two molecules (i.e., to define precisely the concept of biological similarity). Most times this definition will be tailored to the specific aim of a particular system. In this way, Holm and Sander (1996) developed the Dali method, which is a general approach for aligning a pair of proteins represented by 2D matrices. These two-dimensional distance matrices are a representation of a 3D structure. In these matrices the ijth element denotes the distance between the ith and jth atoms. They are useful for the comparison process, since similar 3D structures have similar interresidue distances. So, the problem of matching two protein structures turns into a matrix matching problem. The Dali server can be accessed through the WWW or via e-mail. The user must provide the coordinates of a structure as query input. As a result, a list of similar structures is returned.

In the same context, Shindyalov and Bourne (1998) proposed an algorithm (CE, combinatorial extension of the optimal path) to find an optimal 3D alignment of two polypeptide chains. In order to do so, characteristics of their local geometry (defined by vectors between C alpha positions) are used. One-on-one structure alignment using such an algorithm is available via the Web (http://cl.sdsc.edu/ce.html). Users must submit complete or partial polypeptide chains in PDB format. Then, statistics for the alignment are returned along with the sequence alignment resulting from the structure alignment. Searches against the complete PDB database using a coordinate set not found in the PDB are also possible (afterward, results are mailed to the submitter). In this sense, we can consider the CE algorithm a tool that helps the user in accessing the real content of the PDB database.

Another interesting approach to query in PDB by structural similarity to be mentioned here is that proposed by Artymiuk *et al*. (1992). Their work on 3D searching in databases of small molecules aims at supporting drug and pesticide discovery. The authors describe the use of two algorithms (atom mapping and clique detection) for similarity searching in databases of small 3D molecules. Additional issues related to the efficiency of the implemented method as well as more effective ways of representing the data are discussed. Another work by these authors is the algorithm PROTEP (from "protein topographic exploration program"), which is aimed at providing another approach for structural similarity searches in PDB.

3. OVERVIEW OF QUERY-BY-CONTENT SYSTEMS IN MULTIDIMENSIONAL IMAGE DATABASES

Research in visual information systems flourished in the early part of the decade, with the general idea that in many applications visual data like images and videos convey as much information as alphanu-

meric data and hence should be treated as first class searchable objects by database systems. It was recognized that there are two kinds of information associated with a visual object (image or video): information about the object, called its metadata, and information contained within the object, called visual features. Metadata (such as the name of a protein) is alphanumeric and generally expressible as a schema of a relational or object-oriented database. Visual features, in contrast, are mathematical properties of the image derived by computational processes, typically image processing, computer vision, or computational geometric routines, executed on the visual object. For example, the set of Fourier coefficients of the boundary of an object can serve as a computable descriptor of its shape (Ming-Fang and Hsin-Teng, 1998). On the topic of retrieval, a database system that allows a user to search for objects in terms of the above-mentioned computed properties is said to support "content-based retrieval" for visual information.

Recent image information systems, both research (see *IEEE Computer*, 1995, for a sampling) and commercial (from companies like Virage, IBM, and Excalibur), are leaning more toward a query-by-example paradigm. There are two different styles for providing examples. In the first style, the example is pictorial: the user specifies a query by providing some example image and asking the system to find other images in the database that look "similar" to the example image. Systems of this kind include the Virage Image Engine (Bach *et al.*, 1996) and the NETRA system (Ma, 1997). In the second case the user provides value examples for one or more visual features, something like an image with about 30% green and 40% blue with a grass-like texture in the green part. The values are not provided in English, but are provided by using visual tools to choose colors and texture. An example of this kind of system is the QBIC engine from IBM (Niblack *et al.*, 1993).

When a content-based retrieval system is applied to any specific domain it needs to solve two important problems: which computable features are sufficient to describe all images in the domain and what mathematical function should be used to find a measure of similarity between two objects. In the case of general images, both these issues have been studied extensively. The issue of query-by-content can be then regarded as a clustering application in the context of querying in complex databases. Some popular features used in images are histograms of color distributions of an image, texture features computed from the gray-level concurrence matrix of an image, and moments of strong edges in the image for describing shapes. In most cases, the similarity between images is computed by finding a measure of difference (or distance) between images. A popular way of computing such a distance is to compute the Euclidean distance between corresponding features from two images and then computing a weighted average of these individual distances.

Indeed, much of the developments presented so far have been aimed primarily at 2D images, with little consideration of multidimensional images. A growing field of interest is that of video databases, and much effort is devoted to this area. In contrast, very few examples exist today on query-by-content applications designed for 3D images. Still, one of the most interesting experimental developments is in the biomedical area. A new content-based 3D neuroradiologic image retrieval system recently appeared (Liu *et al.*, 1998). This system is interesting in that it is a good example of a true content-based image retrieval system within a specific 3D-image domain. It deals with a multimedia database containing a number of multimodal (MR/CT) images as well as collateral information related to each image (patient's age, sex, symptom, etc.). Such a system could help medical doctors to confirm diagnoses, as well as for exploring possible treatments by comparing the image with those stored in the medical knowledge databank. As far as we know, no other developments have been reported in biology or biomedicine so far.

4. UNDERSTANDING DATA SET CONTENT: A view from biology on some relevant structural characteristics of 3D images of biological macromolecules

The previous section outlined the basic steps that query-by-content on a 3D image database involve. Clearly, the precise definition of these steps and the way to combine them are, in general, rather difficult. Therefore, and as a way to reduce this complexity, we are going to make extensive usage in this work of *a priori* knowledge related to the concrete application domain to which this study applies. We are considering querying by content only on 3D images containing structural information of biological macromolecules (a subset of BioImage; Carazo and Stelzer, 1999). Indeed, the explicit consideration of the specific application domain in which queries are going to be performed will be used as a strong restriction principle in the definition of visual features and their combination.

The key simplifying consideration is that an initial specification can be made on some of those structural characteristics that will be the most used when a query-by-content is performed in this specific biological context. Certainly, the choice of a given set of query specifications will introduce limitations into

the querying schema, although this is always the price to be paid when tailoring a general problem to a given application domain. Still, the set of specifications that are considered in this first application of query-by-content is a mixture of rather generic features (which will not introduce intrinsic query limitation) with rather specific features. In fact, this choice reflects a strategy aimed at initiating a feedback procedure between developers and users by which this initial set of domain-dependent specifications will be expanded and refined according to the future demand of the community using this new structural database.

There are five biological structural features that are considered in this work: size, shape, "channels," "internal cavities," and "symmetries." Obviously, the first two are very general while the latter three are quite specific. In this work we have considered the following definitions:

Size: We refer to the dimensions in angstroms along the *x*, *y*, and *z* axes of the minimum bounding-box containing the specimen. The size of the specimen acts as a filtering feature, selecting different broad areas of interest.

Shape: This feature is related to the geometrical appearance of the specimen, and it can either be regarded as a global attribute or refer to specific regions of the specimen.

Channels: We refer to low-density areas that traverse sections of the 3D images. In the specification of this feature we will focus our attention on four parameters: the number of channels traversing the specimen, their length, their pathways, and their diameters.

Internal cavities: The definition of internal cavities adopted in this work is that of low-density regions that are totally enclosed within the volume of the specimen. In addition we will consider both their volume and location.

Symmetries: We refer to characteristic point symmetries of the biological specimen that are specified by their appropriate symmetry elements.

The rationale of the initial choice of these five features is simple. The two general features, size and shape, provide the context information. Then, the three specific features, channels, internal cavities, and symmetries, address characteristic properties that have the potential to provide key biological information. Certainly, the choice of these latter three features for our first implementation has been very influenced by the types of specimens we work with, and it is acknowledged that they should be expanded to other features of special importance in other systems in future implementations. Hopefully, this expansion will be a direct consequence of a fruitful interaction with users of the BioImage server of macromolecular 3D images.

5. SYSTEM LAYOUT

5.1. *User-Provided Annotations versus Automatic Computational Features Extraction*

User-provided key word annotation is the traditional way to introduce knowledge concerning a piece of information to be stored in a database. In the context of general image databases, data sets can be annotated manually adding key words, such as the presence of a relevant aspect (i.e., the occurrence of a famous character), that are later used at the query stage to retrieve related images.

However, the exclusive application of a manual annotation approach presents some problems in at least two areas: (1) it implies a large amount of manual effort in producing an annotated image, making this approach impractical for large collections of data sets; (2) differences in the interpretation of the image content could lead to inconsistencies in the key word assignments among different submitters.

As an approach designed to circumvent these problems, most recent query-by-content systems implement an initial stage of automatic feature extraction at the time of submitting a new data set. During this stage, a number of computational descriptors from the images are calculated and then stored as part of the image description.

Nevertheless, the two approaches commented on before, the manual and the automatic approaches, are not necessarily mutually exclusive, and they can be used concurrently in the design of hybrid systems, as is in fact the case for the implementation described in this work. In precise terms, for the work on query-by-content on 3D images in BioImage, a two-stage hybrid approach has been followed: first, at the level of combining the information contained in the textual key words with the structural information itself (these key words refer to general textual information, termed "metadata" in the work by Lindek *et al.,* 1999), and second, at the stage of extracting the five biological features introduced in the previous section. During the stage mentioned, we have made use of both automatic processing tools and information provided interactively by the submitter.

5.2. *On the Computation of Features*

So far, we have been referring to visual features in a rather abstract way. For instance, we have defined "shape" as being "related to the geometrical appearance of the specimen." However, concrete computational approaches are needed when we reach the

level of assigning numerical values to these features. Indeed, a clear differentiation can be made between the high-level features to which a structural meaning can be attached, on the one hand, and the set of lower level primitive features that are used at the core of the system, on the other hand.

In practical terms, any high-level feature is going to be calculated as a combination of lower level features, when we will refer to it below as "computational features." In addition, for abstract features such as shape, the way the combination of computational features is done—or even the set of these features to be used for this task—is not unique. This problem, quite common in image understanding systems, is usually referred to as "the semantic gap." This term expresses in a rather direct way that a gap exists at the semantic level between the abstract means employed by the user to address the understanding of an image and the concrete mathematical functions that the system really computes. As a possible way to partially bridge this gap, we have introduced the notion of a "bio-vocabulary," which is a set of predefined terms that map some of the user abstractions into defined relationships between sets of computational features. We will expand this notion later in this work.

In more precise terms, a computational feature f_i is defined by a tuple $\langle N_i, S_i, D_i, I_i \rangle$, where N_i is the chosen name for the feature, S_i is the space of representation, D_i is a distance function from which a similarity measure is obtained, and I_i is an index structure needed to store and access efficiently the data sets' values for that feature. It should be noted that for a particular feature several spaces, distance functions, and index structures could be chosen. A structural feature is then defined as a combination of computational features satisfying expert criteria related to the information being considered.

In the following paragraphs we will review the five high-level structural features introduced previously, making explicit the way they are calculated in practice through a set of computational features. An important consideration that we must make at the onset is that, within the particular implementation of the prototype presented in this work, most of the computational features are extracted from isosurfaces derived from the 3D images, rather than from the 3D images themselves. Indeed, we believe that this is a topic in which further work is needed in order to extract more information from the 3D density distributions directly. These are the steps to follow:

Size (dimensions). In order to compute the size of the specimen in angstroms, a procedure has been implemented that computes the smallest bounding box containing it. To do that, we first obtain a reference coordinate frame positioned at the center of mass of the object. This frame is derived by aligning the object with the principal axes of the tensor of inertia. Then, the height, width, and depth of such a box are calculated, as well as the occupation rate inside it. Computation of several invariant moments has been done in order to obtain normalized representations for every data set against shift and rotation (Galvez and Canton, 1993).

All the operations that lead to the specification of the coordinate system of size are computed on the 3D images automatically, i.e., without requiring any intervention from the submitter. This operation is done in a short time.

Global shape. As we have commented before, shape is an example of a rather abstract structural feature in which all the difficulties associated with the semantic gap problem are especially relevant. As far as humans are concerned, the traditional way to describe shapes is through the use of words; thus, it is clear what a ring-shaped or a spherical-like object is. However, computer programs need quantitative measures in order to "recognize" and compare such shapes. Thus, several ways to represent the shape of an object have been investigated. One of the most popular ways is the use of histograms as statistical distributions of some function value. In our case, we compute automatically the surface shape spectrum, which is a histogram of a shape index (Nastar, 1997). The shape index is defined as

$$S(p) = 0.5 - (\pi)^{-1} \cdot \arctan[(k1(p) + k2(p))/(k1(p) - k2(p))],$$

where $k1(p)$ and $k2(p)$ are the principal curvatures at a surface point p.

Along with the shape index, the feature vector formed by the moments of inertia can also be used in the process of identifying the different shapes, even though a combination of the shape index and the moments of inertia has not been implemented in this prototype yet.

A bio-vocabulary is introduced at this stage; while it is clear from the previous paragraphs how the computational features associated with shape are computed, it is not clear how they can be combined into an abstract description of shape. This problem is indeed extremely complex and fundamental in image understanding, and instead of attempting to provide generic solutions to fundamental, long-standing, mathematical problems, our aim has been to explore partial solutions with practical application within a defined context. In this way, we have

selected for this prototype implementation a reduced set of abstract shapes, such as ring-like, spherical, and so on, thus setting up our initial bio-vocabulary. For each abstract shape considered in this prototype (toroidal, spherical, triangular), a set of models has been generated. Then, the histogram of shape index has been extracted from them, so as to define a set of multidimensional points—one per model—around which the multidimensional points associated with the experimental 3D images have been clustered. In this way, in addition to having stored in the database the histogram of shape index that is automatically extracted from the 3D images, a "shape label" is also introduced during the clustering process by assigning the label of the closest cluster to each experimental 3D image. This assignment of labels to 3D images creates, in practice, a bio-vocabulary dictionary (see Fig. 2). As will be further noted in the Section 6, this label will be especially important both when querying by specifying complex features in abstract terms and at the resulting visualization interface.

This global shape feature can be complemented with another, a "local shape" feature. The calculation of the latter would follow the same principles as the former, although it is necessary to develop an additional tool dedicated to "cut" some interesting structural regions and which has not been incorporated in this prototype yet.

Symmetry: The symmetry-related features are not calculated automatically. Instead, the submitter must provide information about the type of symmetry and the additional information associated with it. For the time being, only rotational symmetries are considered, and the submitter is required to provide the direction of the symmetry axis in the original reference system in which the 3D image was initially submitted.

Channels (number, path, diameter): Descriptors of channels traversing the specimen can be extracted with a minor user interaction. For example, in order to derive the path of a determined channel, the user must provide a set of seed points along that channel, including its starting and ending points. Then, an algorithm, which calculates the shortest path along these points, is applied. At present, the submitter must provide the number of channels. However, we are considering doing this task automatically by computing the Euler number. The Euler number is a topological invariant measure from which the number of channels and internal cavities can be derived (Lee *et al.*, 1991).

Internal cavities: The user interactively provides one point inside the cavity, around which the volume occupied by such a cavity and its surface histogram are computed with the help of a filling algorithm.

As stated in the description above, some of the features need the user's interaction in order to be properly derived. For instance, local shape and channel characterization are typical examples that demand visual annotation tools to introduce some initial information. At present, this process is carried out by relatively simple tools running on the server side, mainly involving the display of 2D sections, even though virtual reality modeling language (VRML)-based interactive tools (Hartman and Wernecke, 1996) are under development. An example of a possible user input in a VRML world is provided in Fig. 1.

5.3 System Structure

The development of a query-by-content system involves interaction among three main areas, namely, the feature extraction programs, the database technology, and the visual tools interfacing with the user. A diagram representing the flow of events among these components is shown in Fig. 2.

Each data set has its content extracted in a hybrid way combining interactively user-provided information with automatic procedures. The aim of manual annotations is to obtain information that cannot be easily extracted by pure computational methods. Sometimes, the real reason to obtain the information in a nonautomatic manner is the intrinsically more abstract nature of these data. However, in order to make this task easier, user-friendly annotation tools must be provided, thus justifying the further development of 3D environments in combination with complex databases such as BioImage (for general considerations see Pittet *et al.*, 1999).

During the design of automatic feature extraction programs, the trade-off between the required computational time and the precision of the representation for a given visual feature should be considered. In general, representations that lead to efficient and fast algorithms are preferred to more complex and computational time-consuming approaches, even if they might be more accurate. There is a possible risk of obtaining false detections, but the faster system response may compensate for this drawback.

From the database standpoint, both content's descriptions (automatic and textual) need to be indexed and structured in a regular database management system. Actually, the database issue is a critical standpoint since efficient retrieval performance is a key requirement in every query-by-content system. The database management system that we have chosen for the prototype's implementation is the Informix Universal server (which implements an object-relational policy) running on an O2 Silicon Graphics machine.

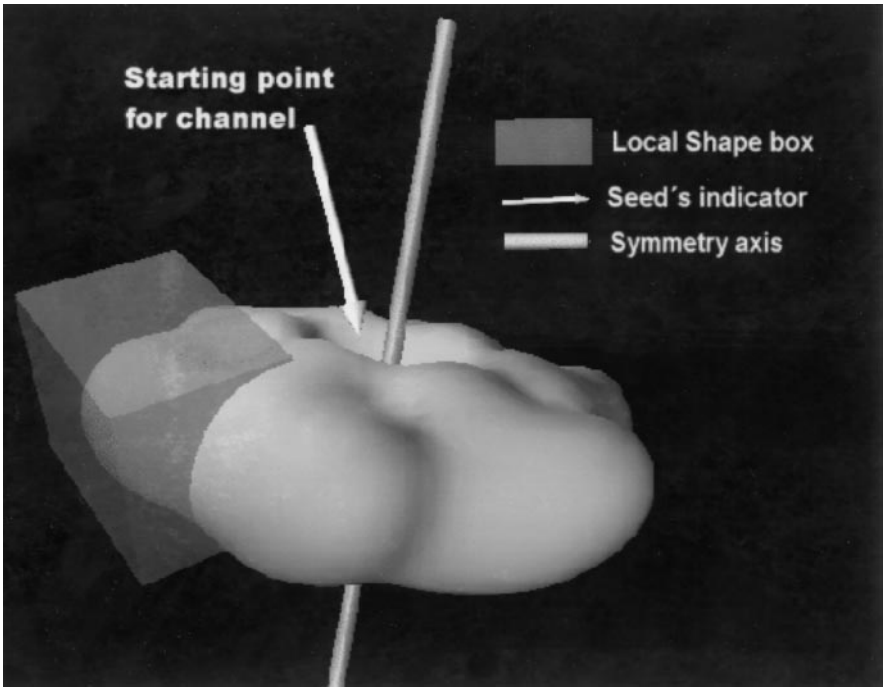

FIG. 1. Example of a VRML world. An annotated isosurface representation corresponding to the data set of the SV40 large T antigen is presented. The central symmetry axis is indicated, as well as the starting point for a channel. A shaded box marks a designated patch on the surface to be further analyzed.

6. QUERY AND VISUALIZATION INTERFACES

The database is accessible via WWW (http://www.bioimage.org/qbc/index.html) as an advanced query tool implemented in the BioImage Web server in Madrid. Similarity queries can be made using two types of query interfaces (see Fig. 3).

The first interface is designed to allow the user to specify properties of the structure to be retrieved (content-related or not). This goal is accomplished by providing a form that is processed by a CGI program. Such a program allows the user to query by means of three kinds of fields: not content related, annotated, and derived from bio-vocabulary. The first type of field includes those normally stored in BioImage as part of its metadata (Lindek *et al.*, 1999). These fields are not directly related to the structural prop-

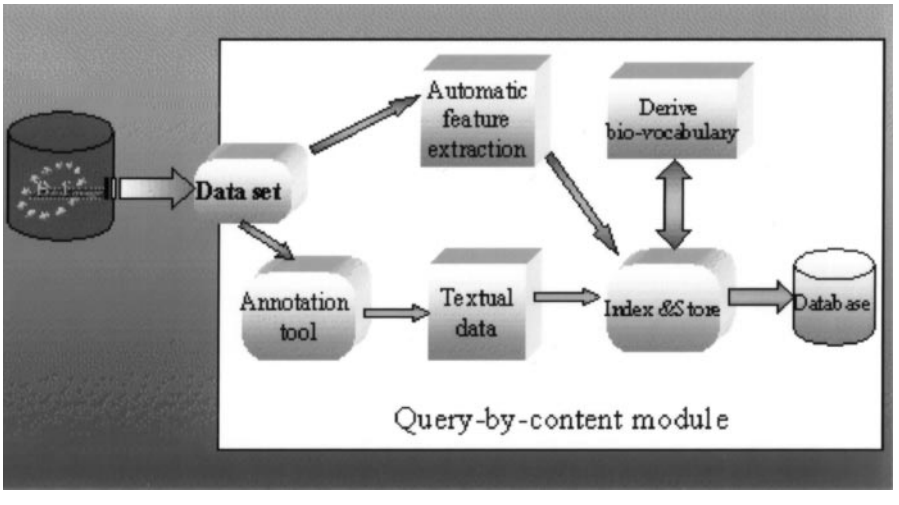

FIG. 2. Components of the query by 3D content engine. BioImage is our source of information. Every macromolecular data set is processed and/or annotated in order to derive information on its content. This information is indexed and stored in a separated subset of the BioImage database. A bio-vocabulary dictionary is also provided by abstracting some of those features.

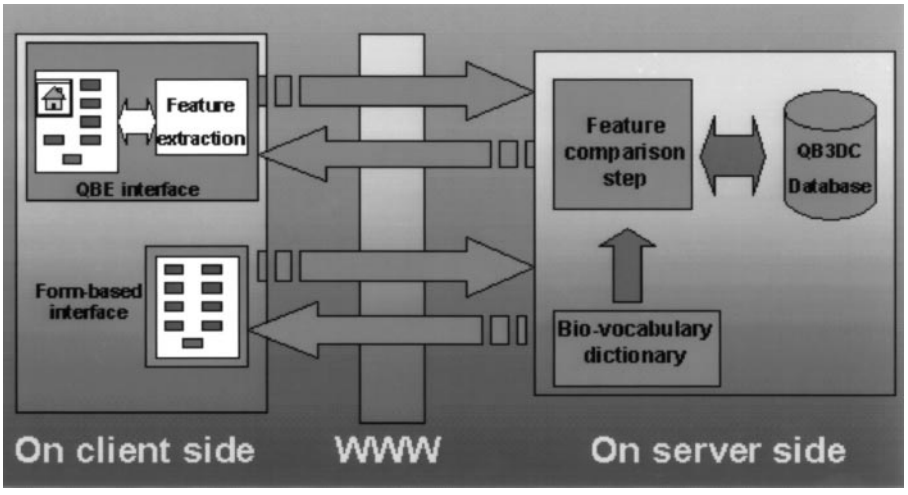

FIG. 3. Interaction between the query interfaces and the search engine. Two query interfaces are indicated at the left-hand side of the image: a form-based interface as well as a query-by-example interface. The form-based query interface allows the user to enter "terms" contained in the bio-vocabulary dictionary as well as other annotated descriptors. The query-by-example interface allows the user to present an example of a structure asking the database to search for similar cases. The latter interface will be complemented with a Java program for feature extraction in the client machine. Results of the query procedure are retrieved in the same way for both interfaces.

erties of the data set but have to do with some other features (e.g., the name of the specimen). The second kind of field is related to those content-based features that are interactively provided by the submitter by a manual annotation procedure, such as the symmetry elements. The third type is formed by those features that are totally or partially computer-extracted. For example, the ring-shape property is stored in the bio-vocabulary dictionary together with the shape spectrum of an "ideal" ring, which is considered as a pattern model. When the user asks for "rings" this pattern model is compared with the spectra of every data set stored in the database. Thus, a distance measure with the label "ring-shaped" is derived. The advantages of this query approach consist of the fact that it is rather easy to use, and in addition, it does not pose any significant network traffic-related problems upon querying.

Along with the form-based interface we have just described, the possibility of a query-by-example interface opens some very interesting possibilities. In essence, a query-by-example interface allows the user to provide a query data set and, as in the case of biological sequence comparisons, ask the database to find similar data sets considering one or several features (shape, size. . .). As the query data set has not been preprocessed before, the database does not contain any description of its features. This latter fact indicates that the content of the query data set must be extracted before similar objects are searched for in the database. This content's extraction will be realized in future implementations on the client side by downloading a Java program and following basically the same procedures as those used during the submission of a new structure.

This approach will also be used in the development of a submission interface that would run on the client side and that would replace the simple data population procedures that have been used in the implementation of this prototype.

As a result of the query, a ranked list of similar objects is displayed on a separate Web page. The user can inspect each of the retrieved data sets. A VRML isosurface is also available so that the user can interactively explore the selected data set.

7. APPLICATION EXAMPLE

The aim of this section is to construct a concrete case of query-by-content on 3D images of biological macromolecules that will illustrate, in a practical way, the usefulness of these new developments. The application presented in this section has been inspired by a recent work of Hingorani and O'Donnel (1998), in which they consider the shape of a number of biological macromolecules in relation to their function as enzymes interacting with DNA. The authors collected the data from several sources, including the PDB database. In the context of this work on query-by-content on structural databases, we highlight their claim that almost all types of enzymes involved in DNA metabolism have a ring-shaped structure or employ one as part of a functional complex. Also, it is interesting to consider that the central channel of the ring is large enough to accommodate either single-stranded or double-stranded DNA.

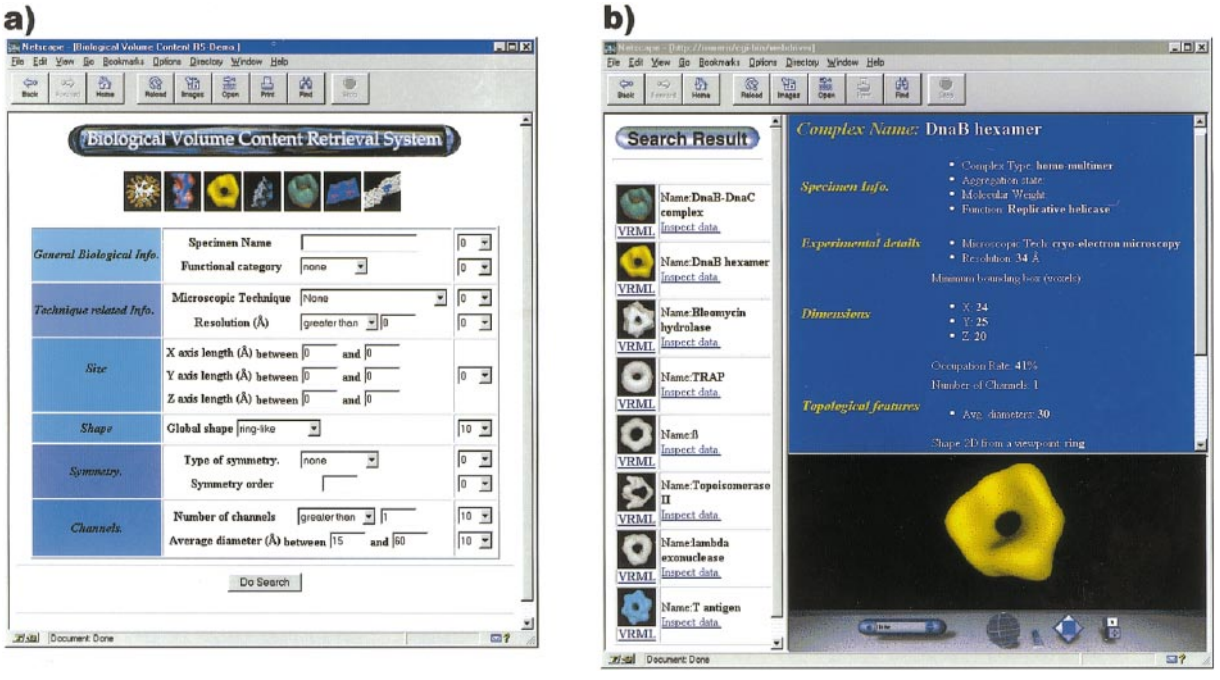

FIG. 4. (a) Form-based interface filled out in order to find 3D images with a ring-like shape as well as an inner channel of a diameter between 15 and 60 Å. (b) Result of the query performed over a set of 3D images of biological macromolecules fullfilling the requirements expressed in the form shown in a. On the left-hand side a list of structures as well as a representative view from each of them is presented. Clicking into any of these structures retrieves general textual information on the specimen and provides a VRML world where that structure can be inspected. In this particular example, further information on the structure of the DnaB helicase, corresponding to the BioImage data set identifier 26, has been selected.

From the standpoint of the database infrastructure development, it is particularly important to consider how their work would have been facilitated should query-by-content approaches over structural databases have already been in common use. Indeed, by searching over the metadata associated with the description of the function, and then retrieving the content information associated with shape, it would have become apparent that all the retrieved structures shared the same ring-like shape.

Taking this work as an inspiration, and with the aim to present a case of a content-based query only, we are presenting as a practical case a search for those structures with a ring-like shape that have a central channel with a diameter of between 15 and 60 Å. The search will be performed over a set of 3D images containing the current BioImage data sets on macromolecules, complemented with some 3D images calculated from PDB files. In this case, the query is form-oriented, and it is performed by filling out the features for shape and channels within the query form (see Fig. 4a). As soon as the user launches the query, the query-by-content engine searches the database for data sets presenting those features. Then, the set of matching data sets is presented on a separate page (see Fig. 4b). The output page allows the user to visualize the content's metadata of each matching data set. Further inspection of the metadata content immediately indicates that the retrieved structures are all involved in DNA metabolism. (It should be clearly stated that the original database question posed in Hingorani and O'Donnel (1998) was that described in the previous paragraph, not the shape-only oriented search described in this work and which is used here only for illustrative reasons. Also, the data set over which the content query is performed is limited.) From this page the user can also inspect an isosurface of the data set that is represented in VRML format.

Another approach to search a structural database focusing on its structural content is that based on a query-by-example interface. Under this approach, we assume that the user has a data set, which is presented as the query structure. In addition, the user must state a set of reference features to be used for searching (e.g., similar shape and an equal number of channels). Afterward, the most similar data sets, in the context of those features, are retrieved. This kind of interface is not yet available, but in the near future it will be one of the key components of our QB3DC (query by 3D content) application.

Users are encouraged to interactively explore the result of using combinations of metadata searches as

well as content-based searches over the five basic structural characteristics implemented in this prototype by accessing the BioImage server in Madrid, where the form-based interface is operative for testing. We believe that this prototype shows clearly the potential benefits of a query-by-content system applied to the biological macromolecule field.

8. CONCLUSIONS AND PERSPECTIVES

In this work we have developed a Web-based prototype system that is able to perform some forms of query-by-content on 3D images of biological macromolecules within the context of the BioImage project. The purpose of this prototype is threefold:

(1) to demonstrate the wide range benefits that the introduction of new search capabilities based on a query-by-content approach can bring to structural databases. In particular, and in the context of this work, we have focused our attention on the macromolecular section of BioImage.

(2) to provide an appropriate modeling framework where new attributes of the data sets' content and their corresponding extraction programs can be tested and incorporated.

(3) to obtain feedback from scientists on the macromolecular field in order to enhance the current prototype.

For a query-by-content approach to be feasible, queries must be reasonably fast. In this respect, all the procedures implemented in this prototype could be scaled up to large ensembles of data sets in simple manners. For instance, the task of extracting some of the visual features could be assigned to the client computer that is submitting the data, rather than in the server (and, in any case, they must be calculated only once per data set). Also, there are many choices of distance functions among volumes, and they can be adjusted for the needed trade-off between speed and accuracy.

The prototype, with its acknowledged limitations, can be therefore considered a pioneer test bed on which more elaborated developments in the field of content data modeling, automatic feature extraction, and interfaces will be tested. So far the focus has not been centered on getting many volumes into this prototype but, rather, in exploring new ways to handle the volume information. Since further work proceeds into query-by-content schema, more efforts will have to be devoted to a powerful and user-friendly submission interface. The availability of this prototype in the macromolecular BioImage Web server is expected to help increase the interactions between developers and users, in such a way that future versions will incorporate new capabilities based on this feedback interaction.

This work has been partly funded by grants from the Comisión Interministerial de Ciencia y Tecnología (BIO97-1485-CE, BIO98-0761) within the context of the BioImage project (funded by the European Union through Grant PL960472). We express our thanks to the participants in the 1997 Workshop on Large Databases of Complex Objects as well as the 1998 Workshop on Query-by-Content and Data Mining in Image Databases, organized with the support of the Spanish Research Council within the context of our joint unit with the University of Malaga. P.A.A. is the recipient of a predoctoral fellowship from the Comunidad de Madrid.

REFERENCES

Artymiuk, P. J., *et al.* (1992) Similarity searching in databases of three-dimensional molecules and macromolecules, *J. Chem. Inf. Comput. Sci.* **32**(6), 617–630.

Bach, J. R., Fuller, C., Gupta, A., Hampapur, A., *et al.* (1996) The Virage image search engine: An open framework for image management, *in* Proceedings on Storage and Retrieval for Still Image and Video Databases IV, San Jose, CA, 1–2 February, Vol. 2670, pp. 76–87.

Carazo, J. M., and Stelzer, E. (1999) *J. Struct. Biol.*, **125,** in press.

Galvez, J. M., and Canton, M. (1993) Normalization and shape recognition of three-dimensional objects by 3D moments, *Pattern Recognition* **26**(5), 667–681.

Gupta, A., and Jain, R. (1997) Visual information retrieval, *Commun. ACM* **40**(5), 70–79.

Hartman, J., and Wernecke, J. (1996) The VRML 2.0 Handbook: Building Moving Worlds on the Web, Addison-Wesley, Reading, MA.

Hingorani, M. M., and O'Donnel, M. (1998) Toroidal proteins: Running rings around DNA, *Curr. Biol.* **8,** R83–R86.

Holm, L., and Sander, C. (1996) Alignment of three-dimensional protein structures: Network server for database searching, *Methods Enzymol.* **266,** 653–662.

IEEE Computer, September 23–31, 1995.

Lee, C., Poston, T., and Rosenfeld, A. (1991) Winding and Euler numbers for 2D and 3D digital images, *CVGIP Graphical Models Image Proc.* **53**(6), 522–537.

Lindek, S., Fritsch, R., Machtynger, J., de Alarcón, P. A., and Chagoyen, M. (1999) Design and realization of an on-line database for multidimensional microscopic images of biological specimens, *J. Struct. Biol.* **125,** 103–111.

Liu, Y., Rothfus, W. E., and Kanade, T. (1998) Content-based 3D neuroradiologic image retrieval: Preliminary results, *in* Proceedings of CAIVD '98: IEEE International Workshop on Content-Based Access of Image and Video Databases.

Lohman, T. M., and Bjornson, K. P. (1996) Mechanisms of helicase-catalyzed DNA unwinding, *Annu. Rev. Biochem.* **65,** 169–214.

Ma, W. Y. (1997) NETRA: A toolbox for navigating large image databases, Ph.D. Dissertation, Department of Electrical and Computer Engineering, University of California at Santa Barbara.

Ma, W. Y., and Manjunath, B. S. (1996) Texture features and learning similarity, *in* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 425–430.

Ming-Fang, W., and Hsin-Teng, S. (1998) Representation of 3D Surfaces by two-variable Fourier descriptors, *IEEE Trans. Pattern Anal. Machine Intell.* **20**(8).

Nastar, C. (1997) The Image Shape Spectrum for Image Retrieval. Research Report 3206. INRIA Rocquencort.

Niblack, W., Barber, R., *et al.* (1993) The QBIC Project: Querying image by content using color, texture and shape, *in* Proceedings of SPIE on Storage and Retrieval for Image and Video Databases, pp. 173–187.

Nucleic Acids Res. January 1999, special issue on biological databases, **27**(1).

Pittet, J. J., Henn, C., Engel, A., and Heymann, J.B. (1999) Visualizing 3D data obtained from microscopy on the Internet, *J. Struct. Biol.* **125,** 123–132.

Rui, Y., Huang, T. S., and Mehrotra, S. (1997) Content-based image retrieval with relevance feedback in MARS, *Proc. IEEE Int. Conf. Image Proc.*

Sedman, J., and Stenlund, A. (1996) The initator protein E1 binds to the bovine papillomavirus origin of replication as a trimeric ring-like structure, *EMBO J.* **15,** 5085–5092.

Shindialov, I. N., and Bourne, P. E. (1998) *Protein Eng.* **11**(9), 739–747.

Smith, D. W. (Ed.) (1994) Biocomputing, Academic Press, San Diego.

Stukenberg, P. T., Studwell-Vaughan, P. S., and O'Donnel, M. (1991) Mechanism of the sliding clamp of DNA polymerase III holeoenzyme, *J. Biol. Chem.* **272,** 11328–11334.

Sussman, Lin, Jiang, Manning, Prilusky, Ritter, and Abola (1998) Protein Data Bank (PDB): database of 3D structural information of biological macromolecules, *Acta Crystallogr. Sect. D* **54,** 1078–1084. [URL: http://www.pdb.bnl.gov/]