Modeling Shape and Topology of Low-Resolution Density Maps of Biological Macromolecules

Pedro A. De-Alarcón,* Alberto Pascual-Montano,* Amarnath Gupta,† and Jose M. Carazo*
*Biocomputing Unit, Centro Nacional de Biotecnologia (CSIC), Campus UAM, Cantoblanco, 28049 Madrid, Spain; and †San Diego Supercomputer Center, University of California at San Diego, La Jolla, California 92093-0505 USA

ABSTRACT In the present work we develop an efficient way of representing the geometry and topology of volumetric datasets of biological structures from medium to low resolution, aiming at storing and querying them in a database framework. We make use of a new vector quantization algorithm to select the points within the macromolecule that best approximate the probability density function of the original volume data. Connectivity among points is obtained with the use of the alpha shapes theory. This novel data representation has a number of interesting characteristics, such as 1) it allows us to automatically segment and quantify a number of important structural features from low-resolution maps, such as cavities and channels, opening the possibility of querying large collections of maps on the basis of these quantitative structural features; 2) it provides a compact representation in terms of size; 3) it contains a subset of three-dimensional points that optimally quantify the densities of medium resolution data; and 4) a general model of the geometry and topology of the macromolecule (as opposite to a spatially unrelated bunch of voxels) is easily obtained by the use of the alpha shapes theory.

INTRODUCTION

The study of three-dimensional (3D) biomolecular structures is central to the understanding of molecular biology. In recent years we have witnessed a steady growth in the number of biomolecular structures resolved by scientists and published for the use of the general scientific community. It is easy to speculate that with rapidly growing interest and funding initiatives in structural genomics (Structural Genomics, special issue of *Nat. Struct. Biol.*, November, 2000), we will soon witness a dramatic increase in the amount of available structural data. While managing large amounts of 3D structural information is a challenge in its own right, the problem is further compounded by the ever-increasing complexity of the structures themselves. It is imperative to develop more sophisticated techniques to analyze, represent, store, and search for complex, 3D structural information.

Among the wide variety of efforts undertaken by different research groups to manage and maintain databases of 3D structures, the Protein Data Bank/Macromolecular Structural Database (PDB/MSD) (Sussman et al., 1998; Keller et al., 1998; Berman et al., 2000) holds a distinctive place. PDB/MSD was designed to store and manage 3D structures of proteins solved at atomic resolution by x-ray crystallography or nuclear magnetic resonance (NMR). The utility of creating such a database is evident from the variety of scientific research it has inspired. Artymiuk (1992), Holm and Sander (1993), Shindialov and Bourne (1998), Westhead et al. (1999) and many others have used the system to investigate structural similarities, biochemical properties, and sequence-derived information. Also, Edelsbrunner et al. (1998a), Norel et al. (1999), Morris et al. (2000) and others have studied ligand-protein interactions including docking analysis, focusing on structural information related to the shape, and geometry of the macromolecules.

From the biological perspective we are interested in the detailed characterization of macromolecular surfaces and topologies. However, rather than atomic resolution data as found in the PDB, we focus on medium resolution data produced by other techniques, in particular cryomicroscopy. Indeed, recently, a new approach combining cryomicroscopy with image processing tools has been introduced among the range of quantitative techniques for the solution of 3D structures of macromolecular complexes (Baumeister and Steven, 2000). The strength of this experimental approach is that it does not require specimens to be arranged into crystal lattices (as is the case for x-ray diffraction at atomic resolution) or to be below a given molecular weight (as is the case for NMR). A shortcoming is that it provides structural information mostly in the resolution range between 0.7 nm and 2.5 nm, reaching atomic resolution only in some exceptional circumstances, as for the recently solved structure of AQP-1 (Murata et al., 2000). It has been recognized (Grimes et al., 1999; Kalko et al., 2000; Bohm et al., 2000; Baumeister and Steven, 2000) that the lower-resolution structures obtained from these techniques nicely complement atomic resolution data. Efforts are underway to fit atomic resolution data from x-ray and NMR into larger structures solved by cryomicroscopy, much like fitting pieces into a large puzzle (Volkmann and Hanein, 1999; Wriggers and Birmanns, 2001). A new initiative, called the IIMS project (integrating information on macromolecular structures), coordinated by the European Bioinformatics Institute (Keller et al., 1998), aims at integrating these complementary pieces from different resolution ranges of information into interrelated databases.

Submitted May 14, 2001 and accepted for publication March 22, 2002.

Address reprint requests to Dr. Jose Maria Carazo, Universidad Autonoma, Centro Nacional de Biotecnologia-CSIC, Madrid 28049, Spain Tel.: 34-91-585-45-43; Fax.: 34-91-585-45-06; E-mail: carazo@cnb.vam.es.

© 2002 by the Biophysical Society

0006-3495/02/08/619/14 $2.00

Nevertheless, there is an important technical challenge at the database level in integrating atomic resolution and lower resolution information: medium resolution data have an intrinsically different representation compared to atomic resolution data. Atomic resolution data represent the precise coordinates of atoms forming the structure. In contrast, datasets at medium resolution are presented as density maps defined over a 3D discrete grid in which each point (voxel) has an associated density value. In addition, because medium resolution data resolve larger structures, the typical size of a medium resolution dataset (the number of voxels) may be quite large, and need more complex data management to ensure efficient storage, access, and manipulation.

The goal of the current paper is to develop an efficient way of representing low- and medium-resolution volumetric data in a database. This novel data representation has a number of interesting characteristics: 1) it provides a compact representation in terms of size; 2) it contains a *subset* of 3D points that optimally approximates the probability density function of the voxel values in a maximum likelihood sense; 3) it makes use of the well-established theory of alpha shapes to represent the geometry and topology of the macromolecule (as opposed to a spatially uncorrelated set of voxels); and 4) it allows automatically segmenting and measuring a number of interesting structural characteristics such as passing channels, cavities, and voids, which can be used for posing interesting queries on substructures of the whole macromolecule.

The proposed representation models the atomic cryomicroscopy data in terms of a new quantitative technique called the kernel c-means proposed in Pascual-Montano et al., 2001. Combining this clustering technique with the theory of alpha shapes (Edelsbrunner and Mucke, 1994), we effectively approximate the overall density distribution of the actual molecule with a relatively small number of pseudo atoms from which an alpha complex is built. Once the alpha complex is defined, we can derive a number of auxiliary metrics that characterize geometry, topology, and connectivity properties of the entire molecule, which in turn can be used to search for other molecules having similar geometric properties.

The rest of the paper is organized as follows. In the next section we describe a mathematical technique to represent a 3D volume using a reduced set of points such that their overall probability density function approximates the density profile of the original volume. We refer to this reduced set of points as the "pseudo-atom set" or simply "pseudo-atoms." Subsequent sections provide a brief introduction to the theory of alpha shapes and introduce an algorithm that starts with 3D density maps from electron microscopy, extracts the complete data structure of pseudo-atoms, and builds the alpha complex. A number of interesting structural characteristics that can be derived from our new pseudo-atom-based representation are also provided. Experimental results with both synthetic and real 3D medium resolution density maps are then shown. We will show how the topology and shape of the original biomolecule are indeed preserved in the proposed representation. The experimental results also demonstrate that the representation facilitates automatic and efficient detection and measurement of structural features such as cavities and channels. Finally, we conclude with a summary of our primary results and present a discussion of the new research avenues opened by this work.

VECTOR QUANTIZATION OF 3D LOW-RESOLUTION DENSITY MAPS

In this section we present a technique to sample the original density map with a faithful approximation, one that uses a reduced number of points but still retains the overall shape of the original density map. This class of techniques, often called vector quantization, has been widely used in the literature (Gersho and Gray, 1992) for dimensionality reduction. In the domain of cryomicroscopy, (Wriggers et al., 1998, 1999) have used a vector quantization technique that preserves the geometric relationship between the original and quantized version of the data for the task of docking atomic resolution structures into medium resolution maps. More generally, any vector quantization technique selects a specific property of the input data that is optimally preserved in the approximate version.

In this paper we propose a novel neural network method based on a cost function explicitly designed to compute a set of representative vectors whose probability density closely resembles the probability density of the input data. Probability density function (pdf) is a fundamental concept in statistics. It gives a natural description of the distribution of a continuous random variable in a given interval. Our choice to optimally preserve the pdf guarantees that the shape (i.e., the spatial distribution) of the alpha complex closely approximates that of the original volume. The following section will briefly describe the method. A more detailed description can be found in Pascual-Montano et al., 2001.

Kernel probability density estimator clustering technique (kernel c-means)

A natural way for quantizing a given data space is to partition the space into groups (or clusters) in such a way that all data points belonging to any group can be replaced by a representative data point (code vector) for that group. The task of the quantization method is to estimate the groups from the input data. One of the most widely used methods is based on a distance criterion where the representative data items are selected in such a way that the distance from each datum to its closest representative item is minimal (intracluster distance) and the distances among groups or representatives of each group are maximal (inter-

cluster distance). In other words, the aim is to find compact and well-separated clusters in the data. While it is generally true that the set of representative data items (called code vectors) produced by vector quantization techniques tend to approximate the probability density function of the input data, our technique is designed to ensure that the difference between the probability density functions of the actual and the approximate versions of the data is explicitly minimized.

Density estimation is the construction of an estimate of the pdf from the observed data. Kernel estimators have been widely studied for density estimation, and abundant literature on this topic is already available (Parzen, 1962; Bezdek, 1981; Silverman, 1986), so they will only be briefly introduced here.

Let $\mathbf{X}_i \in \Re^{p\cdot 1}$, $i = 1 \ldots n$ denote the data items (represented as n real-valued column vectors of dimension p); let $\mathbf{X} \in \Re^{p\cdot 1}$ denote a variable. The kernel probability density estimator can be formally described as:

$$\hat{D}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{X} - \mathbf{X}_i;\, \alpha) \tag{1}$$

where K is the kernel, with $\alpha > 0$ the kernel width parameter that controls the smoothness of the estimated density.

Intuitively, the kernel estimator can be seen as a sum of "bumps" placed at the observations (data). The kernel function K determines the shape of the bumps while the parameter α determines the width. A commonly used kernel function is the Gaussian kernel:

$$K(\mathbf{X}-\mathbf{X}_i;\, \alpha) = \frac{1}{(2\pi\alpha)^{p/2}} \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{2\alpha}\right) \tag{2}$$

The new vector quantization technique

Given n data items of dimension p, $\mathbf{X}_i \in \Re^{p\cdot 1}$, with $i=1. . .n$, the problem is to find c surrogate "representative" data items (code vectors) $\mathbf{V}_j \in \Re^{p\cdot 1}$, with $j=1. . .c$, such that the estimated density:

$$D(\mathbf{X}) = \frac{1}{c}\sum_{j=1}^{c} K(\mathbf{X} - \mathbf{V}_j;\, \alpha) \tag{3}$$

resembles as well as possible the density of the given data. In the present application the data items $\mathbf{X}_i \in \Re^{3\cdot 1}$ represent the voxels of the EM volume under analysis, properly selected to represent their density, and the resulting code vectors $\mathbf{V}_j \in \Re^{3\cdot 1}$ will be the "pseudo-atoms" or representative data items that are going to be used in subsequent steps of the proposed methodology. Note that the dimension p of the vectors is 3 in this application (voxels' coordinates).

In general, let $D(\mathbf{X};\, \theta)$ be the probability density for the random variable $\mathbf{X}$, where θ is some unknown parameter vector. Let $\mathbf{X}_i \in \Re^{p\cdot 1}$, $i=1. . .n$, denote the data items, then:

$$L = \prod_{i=1}^{n} D(\mathbf{X}_i;\, \theta) \tag{4}$$

is the likelihood function, and the most common statistical estimator for θ is obtained by maximizing Eq. 4. Note that in this case the parameter vector is composed by the code vectors $\mathbf{V}_j \in \Re^{p\cdot 1}$ and the kernel width α, i.e., $\theta = \{\{\mathbf{V}_j\},\, \alpha\}$.

From Eq. 3 and 4, the log-likelihood is:

$$l = \sum_{i=1}^{n} \ln(D(\mathbf{X}_i)) = \sum_{i=1}^{n} \ln\left(\frac{1}{c}\sum_{j=1}^{c} K(\mathbf{X}_i - \mathbf{V}_j;\, \alpha)\right) \tag{5}$$

The aim is to find the values of $\mathbf{V}_j$ and α that maximize Eq. 5. This new method attempts to achieve a vector quantization in such a way that the generated code vectors tend to have the same statistical distribution of the original dataset. In other words, the location of the generated code vectors are estimated such that their estimated probability density is as similar as possible to that of the original data, achieving both goals: vector quantization and probability density estimation.

The following algorithm solves the previously stated problem:

1. Given the input data $\mathbf{X}_i \in \Re^{3\cdot 1}$ (input volume voxels), $i = 1. . .n$; and given a number of "pseudo-atoms" c;
2. Initialize $u_{ji} \in \Re^{c\cdot n}$, for $i = 1. . .n$ and $j = 1. . .c$, satisfying the following constraints:

$$\left\{ \begin{array}{ll} u_{ji} > 0, & \forall\, i, j \\ \sum_{j=1}^{c} u_{ji} = 1, & \forall\, i \end{array} \right\}$$

3. For $j = 1. . .c$, compute the new values for the pseudo-atoms $\mathbf{V}_i$ by using the following equation:

$$\mathbf{V}_j = \frac{\sum_{i=1}^{n} \mathbf{X}_i u_{ji}}{\sum_{i=1}^{n} u_{ji}}$$

4. Compute $\hat{\alpha}$ (kernel width) by using the equation:

$$\hat{\alpha} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ji} \|\mathbf{X}_i - \mathbf{V}_j\|^2$$

5. For $i = 1. . .n$ and $j = 1. . .c$, compute u_{ji} by using the equation:

$$u_{ji} = \frac{K(\mathbf{X}_i - \mathbf{V}_j;\, \hat{\alpha})}{\sum_{i=1}^{c} K(\mathbf{X}_i - \mathbf{V}_k;\, \hat{\alpha})}$$

6. Go to step 3 until convergence.

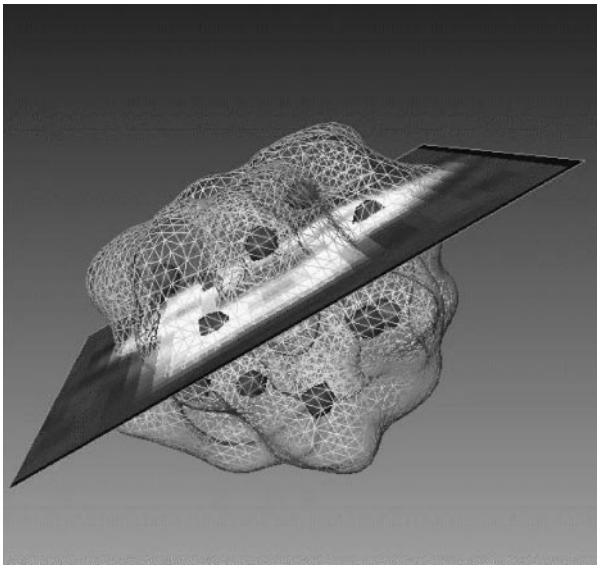

FIGURE 1 Schematic representation of the vector quantization result. In this case, the kernel c-means procedure was requested to produce 12 pseudo-atoms (solid balls) for a 20 Å low-resolution map of bleomycin hydrolase (pdb code: 1GCB). The hexameric structure can be considered as a trimer of dimers. The pseudo-atoms set is easily clustered into three groups (one per dimer) of four pseudo-atoms each.

The algorithm also calculates the average quantization error for each pseudo-atom, which is simply the average distance from the pseudo-atom to the input vectors (voxels) it represents. This measure can be used for evaluating the quality of the quantization, and it is used extensively in other sections of this study. It should be noticed that different kernel functions could be used with this algorithm. In fact, the general multivariate Student's t probability density kernel was used for the experiments carried out in the present work with very good results. It is interesting to note that this approach provides us with an explicit characterization of the pdf calculated from the reduced set of data. As the kernel function family is a choice of the user, the width of the kernel together with the new point set are provided by the algorithm. The information obtained from the process of vector quantization provides part of the data structure that represents the 3D volume, namely the reduced set of data points (pseudo-atoms), and for each of them, their average quantization error. As an example, a schematic view of how 12 pseudo-atoms were placed within the low-resolution map of Gal-6 (Joshua-Tor et al., 1995) is shown in Fig. 1.

ALPHA SHAPES: CHARACTERIZING TOPOLOGY, SHAPE, AND CAVITIES

Until now, we have dealt only with density value distributions and ways to sample the input volume into a reduced point set that approximates the original voxel density pdf. However, geometric properties (shape) are not defined for a set of points, although they are essential to understand a broad range of key characteristics of biological macromolecules. Therefore, a further step to obtain the connectivity between these points (i.e., define their shape) is needed.

A number of issues arise along the process of coding topology and shape-related properties from EM medium resolution datasets that make this task especially complicated. Here we will concentrate on two of them.

First, most approaches that treat geometric information require the definition of some form of molecular surface. There are a number of well-known molecular surface models for structures at atomic levels of resolution (Connolly, 1983a,b; Connolly et al., 1996), primarily because the very nature of high-resolution data lends itself to the definition of theoretically precise surface models. This is not true for medium resolution data, where some segmentation algorithm has to be applied to extract the actual contour of the 3D object, thus introducing an extra level of variability and imprecision.

The second issue concerns the obvious fact that features across different resolution ranges may not be preserved. Important geometric characteristics such as clefts and channels may change their shape and size (or actually disappear) when data at atomic resolution and at medium resolution are compared. Therefore, resolution is a key parameter to be carefully considered when sets of volumetric data are compared. A more detailed study of this general phenomenon, particularized to a selected set of macromolecular features, is underway.

We start by first segmenting the contour that delineates the actual 3D object from its background by the use of Deriche filtering (Deriche, 1987). Deriche's filter is a variation of Canny's filter (Canny, 1986) that lessens its heavy computational load by using a fast recursive implementation. As Canny's filter does, it extracts the contour of the image from gradient information along the three axes (x, y, z). This voxel-based description of the contour of the original object will be taken as the reference for shape.

At this stage of the analysis we have, on one hand, both the original volume of density voxel values and its 3D contour, and on the other hand, a collection of points coming from the density pdf sampling of the original volumetric data. Then, it is necessary to construct a 3D shape using the set of points derived from the pdf sampling. We use alpha shapes theory (Edelsbrunner and Mucke, 1994), a generalization of the convex hull of a point set, to address this issue. Using this theory, it is possible to associate a family of shapes to a finite point set in an n-dimensional Euclidean space. Each shape is a well-defined polytope (an n-dimensional solid with flat faces) derived from the Delaunay triangulation of the point set, with a parameter $\alpha \in \Re$ controlling the desired level of detail.

Alpha shapes theory emerged as a powerful geometric concept that has been successfully applied to the structural biology field—always for datasets at atomic resolution—by

a number of authors (Edelsbrunner et al., 1995, 1998b; Peters et al., 1996). The formal definition of shape and topology provided by alpha shapes makes possible, among other tasks, to efficiently analyze different molecular surface models, and to precisely locate and measure the volume of cavities, voids, and channels (Edelsbrunner et al., 1995). In the following paragraphs we will provide a brief summary of the alpha shapes theory. For further study, we refer the reader to (Edelsbrunner and Mucke, 1994; Edelsbrunner et al., 1995; 1998b, and Akkiraju et al., 1996.

Given a weighted point $P = (p, w_p)$ where $p \in \Re^n$, the *power distance* from a point $x \in \Re^n$ to P is defined as $\Pi_{Product} = \| p - x\|^2 - w_p, \|p - x\|^2$ being the Euclidean distance between p and x. Given a set $\{P_i\}$ of weighted points, the *power diagram* is the tessellation of the space into convex regions (cells) where the ith cell is the set of points nearest to a vertex P_i (in power distance metric). The *weighted Delaunay triangulation* is the face adjacency graph (dual) of the power diagram. Vertices in the triangulation are connected if and only if their corresponding power diagram cells have a common face. The Delaunay triangulation of a point set defines its convex hull. It is composed of a set of k-simplices (0 corresponds to points, 1 to edges, 2 to triangles, and 3 is associated with tetrahedra).

These concepts are not new and have been extensively used in structural biology to derive measures for several surface models (Connolly, 1983b; Connolly et al., 1996). They all build a topological structure based on the regular triangulation of atoms as weighted points. Every atom is considered as a ball $B(p, r_{vw}) \in \Re^3$ characterized by its position (center) and van der Waals radius (weight). Indeed, the weighted Delaunay triangulation defined from the set of atoms of a given molecule provides us with its underlying topological structure (connectivity among atoms).

The alpha shapes theory extends all these ideas by introducing a new growing parameter α. The reader should not be confused with the α parameter used in the vector quantization algorithm. Let's suppose that all the atoms (balls) of this molecule start to grow simultaneously by increasing their radii by the α parameter. Then, every atom will be redefined as a ball $B_\alpha=(p, r_\alpha)$ with radius $r_\alpha = \sqrt{r^2_{vw} + \alpha^2}$. As α increases (see Fig. 2) the balls gradually grow so that they will eventually overlap. At the moment when the boundaries of two balls overlap, a new 1D simplex is added (an edge in this case) to the simplicial complex for that particular value of α. Whenever three balls overlap, a triangle (2D simplex) is created, while a tetrahedron (3D simplex) is added for the case of four intersecting balls. The simplicial complex associated with an α value is a subcomplex of the Delaunay complex and it is called the *alpha complex*. The *alpha shape* is the part of space occupied by simplices in the alpha complex. When $\alpha = 0$ (zero-shape) the actual topological structure of the molecule is obtained; however, when α tends to ∞ the alpha complex is the convex hull of the initial set. Simplices in an alpha complex for the alpha value α_1 also belong to that one for the alpha value α_2 with $\alpha_1 < \alpha_2$. That is, the alpha complex for a smaller value of alpha is always a subcomplex of the one for a larger value of alpha, and both are subcomplexes of the Delaunay complex (convex hull).

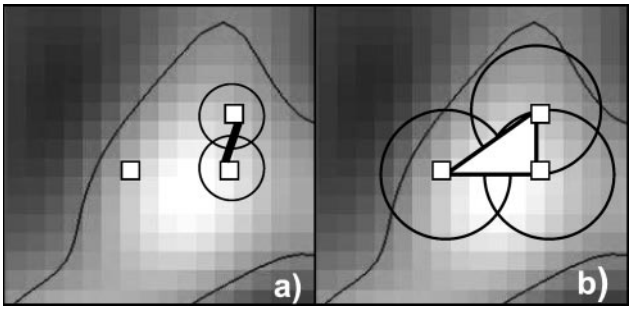

FIGURE 2 Construction of the alpha shape family. (*a*) A new simplex, an edge in this case, is added to the alpha complex whenever two spheres overlap due to an increment of alpha. (*b*) A triangle appears when the alpha value is increased so that three spheres overlap.

Relating the alpha complex to the weighted Voronoi decomposition of the molecule does the link with the actual molecule. The alpha complex encodes combinatorial, topological, and metric information about the molecule.

In this work alpha shapes theory will be applied to medium resolution density maps for which information about the atomic coordinates is not available. The basis of our usage of alpha shapes resides in the prior vector quantization procedure, which has provided us with the pdf-based sampling (pseudo-atoms set) from the original 3D image. Our method will proceed by using these pseudo-atoms in a way similar to the way real atomic positions have been used in the context of atomic resolution data. The initial value of the alpha parameter will be taken as the maximum of the quantization errors associated to each pseudo-atom, as described in detail in the next section. In this way, an alpha complex will be obtained as a representation that encodes both the geometric and topological properties of the medium resolution information.

Once the alpha complex has been obtained, the selection of those k-simplices ($k = 0. . .3$) that form voids and cavities of any number of mouths is straightforward, allowing the automatic detection and quantification of features such as deep clefts and channels. Other interesting structural features are also possible to detect (e.g., protrusions), although they will be further investigated in future work.

ALPHA SHAPES AND 3D TEM DENSITY MAPS: MODELING GEOMETRY AND TOPOLOGY

The general schema to derive our proposed volume data representation is as follows:

1. From the original volume, obtain a point set of "pseudo-atoms" that quantify the volume density. Associated to each pseudo-atom there is a "quantization error." The number of pseudo-atoms is initially arbitrary;
2. Construct the alpha complex associated to the set of pseudo-atoms, use the largest quantization error as the value for the alpha parameter;
3. Increase the number of initially computed pseudo-atoms such that the new alpha complex constructed on them captures the geometrical and topological features of the original volume in a well-defined manner and within a precise and user-controlled error limit;
4. Store the set of final pseudo-atoms and its associated alpha complex. They will constitute the data structure onto which all subsequent density, geometry, and topology analysis will be performed.

This schema presents a number of algorithmic and mathematical issues. To start with, it is centerpiece the notion of connectivity among sampled points (pseudo-atoms). However, one of the alpha complexes must be selected among those that compose the whole alpha family. The alpha family is the set of all possible alpha complexes of the point set indexed by the alpha parameter. Thus, the selected alpha complex should be chosen such that it best preserves the topological and geometrical characteristics of the original macromolecule volume.

Fortunately, the information produced by the vector quantization procedure is not only the set of pseudo-atoms referred to so far, but also a measure of the vector quantization error associated to each pseudo-atom. The quantization error is the average distance from the pseudo-atom to the input vectors (original voxels) it represents. The maximum of such errors provide us with a good, yet conservative, value for the alpha parameter.

Additionally, to compute certain shape measures from the alpha complex it is important to provide the capability to obtain a volumetric object from the alpha complex; in other words, to produce a voxel-based volume from the list of simplices that composes a particular alpha complex. In this way, for each alpha complex it is possible to build a discrete binary image in $\Re^3$ that corresponds to the space occupied by its simplices. That is, every simplex is "discretized" into a set of voxels. A given voxel in the output volume will be assigned 1 as long as it belongs to any vertex, line, triangle, or tetrahedron of the alpha complex (and 0 otherwise). Thus, tetrahedra will approximate the inner mass of the molecule and outer triangles will produce its contour. We keep track of the real molecular dimensions as we know its spacing; i.e., the number of Angstroms that corresponds to each voxel.

The overall goodness—in terms of topological and geometric fidelity to the original dataset—of the information contained in the data structure will depend on two key parameters: 1) the cardinality of the pseudo-atoms set. As it will be shown, the higher the number is, the better accuracy is obtained in the voxel-based model rendered from the alpha complex. 2) the particular choice of the α value for a fixed number of pseudo-atoms. It must guarantee that the 3D shape and topological structure obtained from the simplices of the alpha complex are correct.

To reduce algorithmic complexity, we will fix the α value equal to the maximum quantization error in the calculation of the pseudo-atoms. Therefore, only the cardinality of the pseudo-atoms set will be regarded as a free parameter to be optimized. The optimization is carried out such that the resulting data representation should simultaneously fulfill the following two constraints.

Constraints of the data representation

Topology preservation

In our case, the preservation of topology implies that the three Betti numbers β_0, β_1, β_2 are equal compared to the original volume. One single connected component (β_0) is obtained, and there exist as many cavities, tunnels (β_1), and voids (β_2) as in the original dataset. These two conditions are calculated in a direct manner from the alpha complex (Delfinado and Edelsbrunner, 1995). If topology is not preserved a new point set with a larger number of pseudo-atoms is created.

Shape preservation

A number of shape measures are calculated for the original dataset and the voxelized model obtained from the alpha complex. The similarity between the original dataset and the model for each of such measures must be greater than a given threshold. We have chosen from the literature up to six 3D shape features. These features are computed for both the original and the alpha shape-derived model. The rationale of this choice is to use different shape features as different ways of "viewing" the geometry of the molecule. Euclidean distance is used to compare them. All shape features but the last one (histogram of normals) are computed from the voxelized version of the alpha complex.

The selected shape measures are classified as either boundary-based or region-based features, depending on whether the boundary or the area inside the boundary is coded. For the computation of some of these features it is necessary to define a new reference frame that will be centered at the object's centroid. Thus, we need to compute first the geometric center of the object and its principal axes of inertia. All these features operate on the spatial domain.

Region-based measures

Ratio among principal axes of inertia (Galvez and Canton, 1993; Lohmann, 1998). Let (λ_0, λ_1, λ_2) be the three eigen-

values sorted by value, so that λ_0 correspond to the largest eigenvector and λ_2 to the smallest. The ratio among them gives an idea of the flatness, elongation, or roundness of the object. For instance, if $\lambda_0 = \lambda_1$ and the ratio between λ_2 and λ_0 is large, then the object is mainly flat. If $\lambda_1 \approx \lambda_2$ and the ratio between λ_0 and λ_1 is large, then the object is elongated. If the three values are approximately equal, the object is mainly rounded. It is easy to see that this measure is invariant to rotations and translations (Lohman, 1998).

Cross-correlation. For binary-valued objects (in our case after the segmentation process required to obtain the contour), the cross-correlation between two objects is defined as the normalized difference between their areas once they have been rotated/translated in accordance with their principal axes.

Size. Length, width, and depth of the minimum bounding box containing the object along its principal axes multiplied by the spacing of the dataset samples.

Circular distribution of mass. This feature was successfully used in Ankerst et al. (1999) with high-resolution data. A shape histogram is built from a partitioning of the 3D space into concentric shells and radial sectors that emerge from the center point of the model. The histogram reflects the number of points that fall within a shell or sector.

Boundary-based features

Image shape spectrum (Nastar, 1997). A histogram is derived from the shape index $S(p)$. This function incorporates differential information of the object's contour. It is defined as:

$$S(p) = 0.5 - \frac{1}{\pi} \cdot \arctan\left[\frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)}\right]$$

where $k_1(p)$ and $k_2(p)$ are the principal curvature values at the contour point p. It is not defined for those surface points for which the principal curvature values are equal (e.g., in a "flat" area). It is invariant against rotations and translations.

Histogram of normals (Paquet and Rioux, 1998). The set of 2-simplices that compose the outer fringe of the object forms a triangulation of the surface. For each of these triangles an orthogonal vector is obtained (normal). Then, the angle between the triangle's normal and the two first principal axes is mapped into the form of histogram. Thus, this measure does not consider local similarity and can be very sensitive when two objects present a good overall similarity except for a given part of one of them.

ALGORITHM TO OBTAIN THE PROPOSED DATA REPRESENTATION

With all the previous considerations, now we are ready to introduce the actual algorithm used to derive the proposed data representation from a 3D density image V:

A. Obtain the binary-valued mask $\boldsymbol{MK}$ from the original density map $\boldsymbol{V}$.
B. Initialize $\boldsymbol{n} = n_0$ as the initial number of pseudo-atoms.
   1. Vector quantization: run the kernel c-means program (Pascual-Montano et al., 2001) to select the $\boldsymbol{n}$ pseudo-atoms contained within $\boldsymbol{MK}$ that best quantify the density of $\boldsymbol{V}$ according to the maximum likelihood estimation of its probability density function. The set of n 3D points is called $\{\boldsymbol{S}_i\}$.
   2. Selection of the value for the α parameter as the maximum of the quantization error.
      a. Obtain the alpha complex $\boldsymbol{A}_\alpha = \boldsymbol{C} \cup \{\boldsymbol{S}_i\}$. Where $\boldsymbol{C}$ is the set of k simplices ($k = 1, 2, 3$) (connectivity).
      b. Take those points closest to the contour and translate them to make them to fall onto the molecule's contour (see comment on this procedure below). This action will generate a new set of points $\{\boldsymbol{S}_i\}'$. Obtain $\mathrm{A}'_\alpha = \boldsymbol{C} \cup \{\boldsymbol{S}_i\}'$.
      c. Create from $\boldsymbol{A}'_\alpha$ a voxel-based model $\boldsymbol{M}$ in the 3D Euclidean space. $\boldsymbol{M} = \boldsymbol{M}_{\text{inner}} \cup \boldsymbol{M}_{\text{contour}}$. $\boldsymbol{M}_{\text{inner}}$ correspond to those voxels generated from 3-simplices (tetrahedra) of the alpha complex. $\boldsymbol{M}_{\text{contour}}$ is obtained from the 2-simplices (triangles) placed at the outside fringe of the alpha complex.
   3. Check the accuracy of $\boldsymbol{A}$:
      a. Preservation of the topology: topology (Betti numbers) of $\boldsymbol{A}$ and $\boldsymbol{MK}$ should be the same, otherwise increase n and go to step B1.
      b. Preservation of the shape: calculate the six similarity shape signatures (defined above) and perform an average over their similarity scores with the original dataset. This measure will be used to check the resemblance sim(**M, MK, n**, α) of the model with the original molecule. If sim(M, MK, n, α) $<$ threshold (threshold is provided by the user) then increase $\boldsymbol{n}$ and go to B1. Otherwise go to step C (topology is preserved and geometry resembles the original one at the specified degree of precision).
      c. Store $\mathbf{A}_\alpha$ as the approximating model for the original molecule. Also store "shifted" atoms of $\{\boldsymbol{S}_i\}'$ to render the alpha complex whenever requested. All the shape features previously calculated are also stored and indexed to increase the performance of shape-based retrieval operations.

Two additional remarks must be made. First, the initialization of the number of pseudo-atoms is certainly an important factor in terms of overall efficiency. The kernel c-means procedure is the most time-consuming task, and a good initial estimate is highly recommended. Our particular choice is to start with 10% of the points of the original dataset, with a step increment of 100 pseudo-atoms. This decision does not affect the quality of the final result, but it

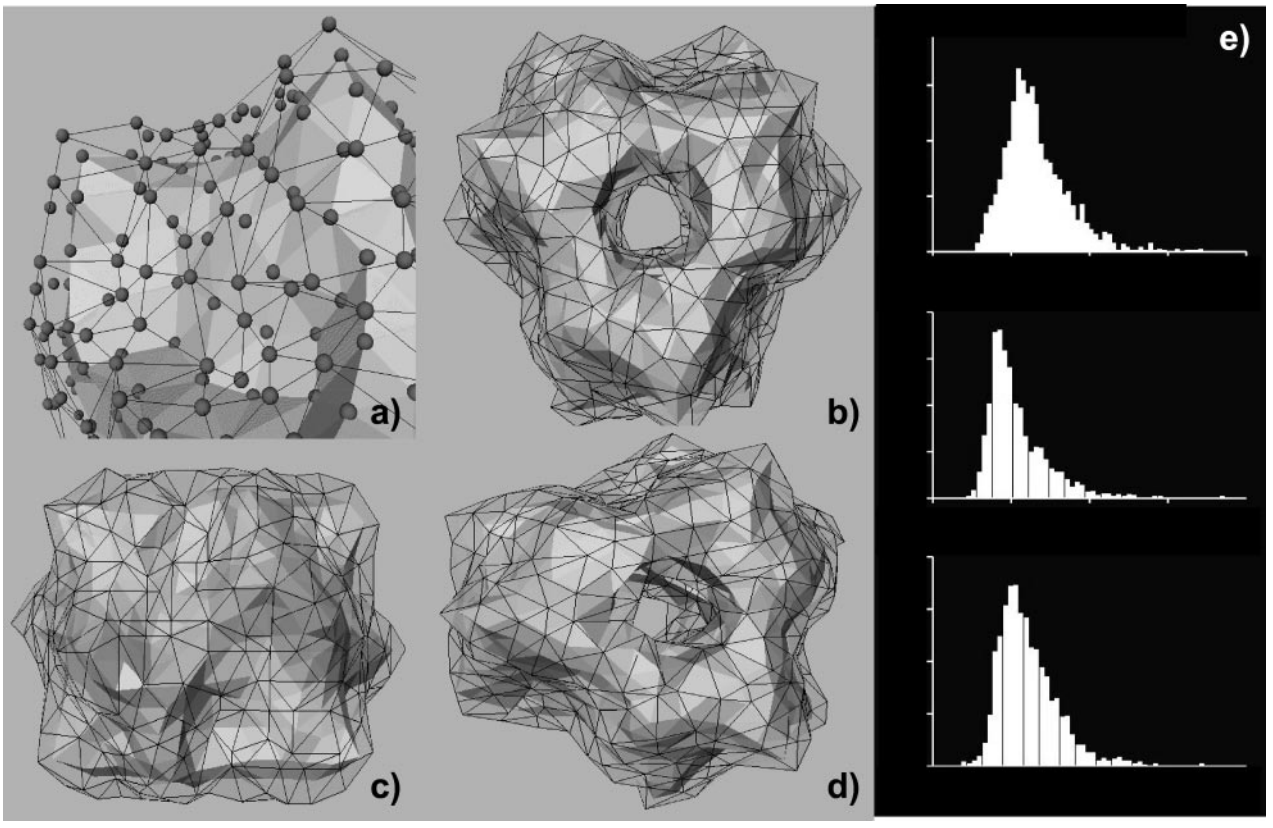

FIGURE 3 Shifting of outer pseudo-atoms. The displacement of outer pseudo-atoms improves the visualization of the simplicial complex whenever a rendering of it is required. (*a*) A detailed view of the "stretched" surface (wireframe) versus the alpha complex's contour (solid) is shown. Different views of the two surfaces are shown in *b–d*. This heuristic procedure is only used for visualization purposes. It does not significantly affect the global or the local shape of the protein. Histograms of normals in *e* describe the global shape of the alpha complex (*top*), the stretched model (*middle*), and the original molecule (*bottom*).

dramatically improves the speed of the overall procedure. Other heuristics are open to particular choices.

Second, in step B2b some pseudo-atoms were automatically "shifted" from their original positions. The rationale for this approach is that because the pseudo-atoms were selected to convey general density information, those pseudo-atoms that define the contour of the alpha complex will not exactly coincide with the real object's contour. Thus, the voxel-based volume rendered from the alpha complex is somehow "scaled down" with respect to the original dataset. The visual effect is that the contour of the alpha complex appears to be located underneath the actual molecular surface. Step B2b was designed as a heuristic approach that compensates the above-exposed effect whenever a rendering of the alpha complex is requested. It is important to note that our representation model is aimed at preserving geometric and topological features. In contrast, visual fidelity is a criterion of interest not for internal representation, but for visualization. In the following we provide a precise specification of this procedure designed to improve visual fidelity when requested.

First the subset $\boldsymbol{S}_{\text{contour}} \subseteq \boldsymbol{S}$ of pseudo-atoms that defines the contour of the alpha complex is automatically extracted (by the use of the alpha shapes tools). Then, for each pseudo-atom contained in $\boldsymbol{S}_{\text{contour}}$, we find out which is its nearest neighbor (following the point's normal direction) located at the actual molecule's contour. In this way, the new set $\boldsymbol{S}'_{\text{contour}}$ is obtained with the same cardinality of $\boldsymbol{S}_{\text{contour}}$. Connectivity relationships (topology) defined by the alpha complex remain unaltered. Moreover, the shape of the alpha complex as represented by the six similarity measures described previously is not affected at the global or local level, as can be appreciated in Fig. 3. Whenever a visualization of the alpha complex is requested, we render its simplices by using the pseudo-atoms set $\boldsymbol{S}' = (\boldsymbol{S} - \boldsymbol{S}_{\text{contour}}) \cup \boldsymbol{S}'_{\text{contour}}$. As a consequence, the set $\boldsymbol{S}'_{\text{contour}}$ needs to be stored in addition to $\mathbf{S}$.

Structural characteristics directly derivable from the alpha complex

The use of the alpha complex provides a powerful means to study a number of interesting structural features of the object. Using alpha shapes in conjunction with flow theory (Edelsbrunner et al., 1998b) it is possible to directly inspect

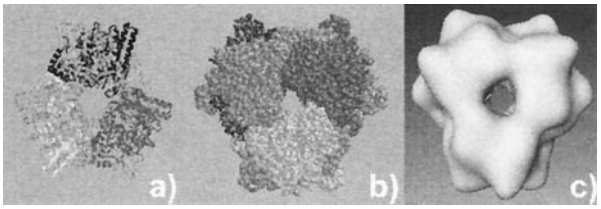

FIGURE 4 (*a*) Secondary structure elements (ribbons) forming the atomic structure of Gal-6 bleomycin hydrolase. (*b*) van der Waals surface of Gal-6. It is the surface of what is covered by the atoms, each represented by a spherical ball with its van der Waals radius. (*c*) Rendering of the surface's Gal-6 map at 20 Å resolution. As the resolution is much lower than for (*a*) and (*b*), most of the surface details are lost. However, the overall shape and the traversing channel are still preserved. Our data representation will capture most of the geometric features of *c*, but in a more efficient and compact manner.

channels of any number of branches, pockets, and voids in the structures. In this way an automatic detection of them is possible, even specifying arguments related to the total volume or the area of the mouths, and then segment them out of the structure for further processing.

RESULTS

Automatic detection and quantification of cavities and channels

The experimental results that we present in this section are aimed at showing that 1) the proposed data representation can be practically obtained with the algorithm presented in the previous section. Also, the representation is robust in the sense that several executions of the algorithm produce almost identical results; 2) this data representation keeps most of the density, topological, and geometric information of the original dataset in a compact manner; and 3) the information is now expressed in a way that allows for an efficient manipulation when inquiring for density, topology, and geometrical characteristics.

Without loss of generality, we will focus our attention on a concrete problem: the automatic detection and measurement of cavities and channels in macromolecules from volumetric datasets representing complex macromolecular structures at medium resolution. This issue is key in our work because it is the only way to accomplish large-scale objective studies. However, most of the tools that exist today for the analysis of medium- or low-resolution structures are manual. Two different datasets have been used in along this work: 1) a low-resolution volume density generated from the atomic resolution data of the bleomycin hydrolase yeast analog Gal-6, and (2) an experimental low-resolution map of the DnaBC complex obtained by cryo-electron microscopy.

The first experiment reported here is aimed at demonstrating the robustness and stability of the methodology. Given a particular dataset, several executions of the algorithm must produce almost identical results. This means that for every execution of the algorithm the positioning of the pseudo-atoms within the volume is stable and the shape, as represented by the alpha complex, is very similar among several executions.

Gal-6 is the yeast homolog of bleomycin hydrolase (Joshua-Tor et al., 1995), a cysteine protease that hydrolyzes the anticancer drug bleomycin. The structure of Gal-6 was solved by x-ray crystallography at 2.2 Å of resolution (Fig. 4). Gal-6 presents a hexameric structure with a prominent central channel. The hexamer has overall dimensions of 125 × 115 × 85 Å. Because of the extensive dimer interaction, the hexamer may be considered as a trimer of

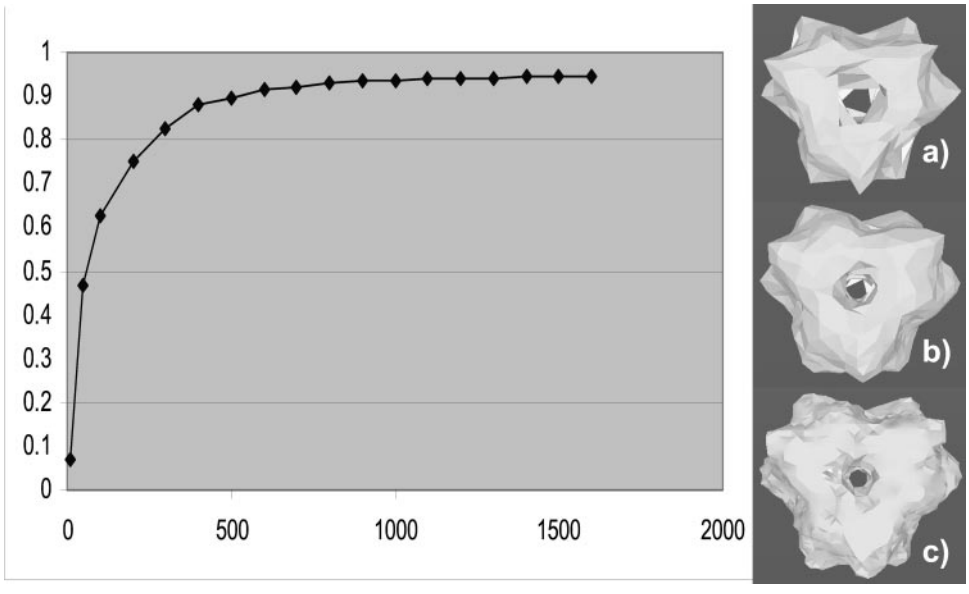

FIGURE 5 Evolution of the similarity between the volume of Gal-6 filtered down at 20 Å and the alpha shape representation. The chart on the left side represents the cross-correlation coefficient (*y*-axis) of the original volume of Gal-6 versus the alpha shape models with increasing number of pseudo-atoms (*x*-axis). The first model with an average shape similarity higher than 95% is obtained with 1500 pseudo-atoms. On the right, the evolution of the alpha complexes with (*a*) 600 and (*b*) 1500 pseudo-atoms is shown. The alpha complex derived from the original surface of the low-resolution Gal-6 volume is presented in (*c*). Note the visual similarity between (*b*) and (*c*).

TABLE 1 Shape stability with different initializations of vector quantization

Feature	Iss	NormX	NormY	Shell	Sector
Cross-difference	$0.0142 \pm 4 \cdot 10^{-6}$	$0.035 \pm 3 \cdot 10^{-5}$	$0.03 \pm 7 \cdot 10^{-6}$	$0.004 \pm 5 \cdot 10^{-5}$	$0.003 \pm 7 \cdot 10^{-7}$
Sim. to original model	$0.951 \pm 7 \cdot 10^{-6}$	$0.928 \pm 8 \cdot 10^{-6}$	$0.937 \pm 9 \cdot 10^{-6}$	$0.986 \pm 3 \cdot 10^{-6}$	$0.99 \pm 6 \cdot 10^{-7}$

Statistics obtained from the analysis of the histogram-based shape features used in this work when calculated on the alpha complex representation of the Gal-6 map at 20 Å resolution. Our algorithm was applied 10 times to the same dataset to check the stability and robustness of the resulting representation. The top row shows the average L2 distance and variance obtained from the "all to all" comparisons. The bottom row shows, for each feature, the average similarity (normalized between 0 and 1) and variance of all the histogram-based shape features calculated between the derived data representation and the original dataset.

dimers. The papain-like active sites are situated within the central channel. The size and shape of this channel, together with the prominent positive electrostatic potential inside the channel, suggest that it represents the region of Gal-6 involved in DNA binding. Also, Gal-6 may undergo conformational changes upon DNA binding that would better allow accommodation of a DNA helix or hairpin in the channel. The atomic resolution data of bleomycin hydrolase Gal-6 were retrieved from the PDB and a volumetric map at 20-Å resolution with a spacing of 3.409 Å/pixel was generated. The CCP4 suite of programs (Collaborative Computational Project, Number 4, 1994. The CCP4 Suite: Programs for Protein Crystallography. *Acta Cryst allogr*. D. 50:760–763) was used for this task.

Following this filtering step, an accurate alpha shape representation of the volume was obtained by executing the algorithm presented above. The similarity threshold used within the algorithm requires that the new data representation had an average shape similarity with respect to the original volume above 95% (as measured by a combination of six shape features). We focus in the behavior of one feature, namely the cross-correlation coefficient (CCC) between the original volume and the voxelization of the alpha complex. It can be readily appreciated from Fig. 5 that the CCC follows a logarithmic-type curve, in which large changes of CCC happen when the number of pseudo-atoms is small, and then the changes become stable. Following the choice of threshold value specified above, the final number of pseudo-atoms required for such an accurate representation is 1500.

Stability of the vector quantization technique

Like many other vector quantization methods, algorithms of the type proposed here are known to be sensitive to initial conditions: the local extreme to which the algorithm converges depends on the choice of initial values for $\boldsymbol{V}$ (pseudo-atoms). In general, the average variability will depend on the shape of the 3D input density distribution and the number of vectors. To test the stability of the algorithm under different initialization conditions, we repeated the algorithm 10 times using different random initialization values for the code vectors. For the tests we used the bleomycin hydrolase dataset at 20 Å resolution with a spacing of 3.409 Å. To check the stability effect of an extreme case, we fitted 1500 code vectors and calculated the RMSE (root-mean-square error) for 10 statistically independent calculations. The stability test showed that the overall RMSE fluctuation of the code vectors is 1.79 Å. This result indicates that the position of the code vectors is very stable. We should emphasize that the vector quantization technique used in this study aims at preserving the pdf of the original data, which means that the algorithm will tend to position more code vectors in those zones with high-density values and fewer code vectors in zones with low-density values, so the separation of the code vectors can be very small in high-density areas of the volume. This fact explains why the stability of the code vector's positions alone is not sufficient to demonstrate the stability of the method. For that reason we also tested the stability of the model (estimated probability density) by calculating the log-likelihood of the estimated density for each independent run, obtaining a mean of 233970.8 and a standard deviation of 4.18, which indicates a very stable model.

In addition to the above tests, we also demonstrated that the shape features coded in the alpha complex also remain stable along different initializations of the pseudo-atom generation algorithm. To this end, the alpha-shape $\boldsymbol{A}_a^i$ was built for each of the pseudo-atom sets ($i = 1...10$). Then, shape features introduced above were calculated for every

TABLE 2 Shape stability with different initializations of vector quantization (2)

Feature	Size (Vox) (z, y, x)	Cross-Corr	λ_0	λ_1	λ_2
Alpha complex (1500 pseudo atoms)	22 ± 0, 33 ± 0, 38 ± 0	$0.98 \pm 3 \cdot 10^{-6}$	$0.829 \pm 1 \cdot 10^{-5}$	$0.837 \pm 4 \cdot 10^{-6}$	1 ± 0
Original model	22, 33, 38	$0.94 \pm 1 \cdot 10^{-6}$	0.832	0.838	1

Statistics obtained from the analysis of the values of "size," "cross-correlation coefficient," and "eigenvalues" calculated on the alpha complex representation of the Gal-6 map at 20 Å resolution. Top row, average and variance of the values calculated from 10 different data representations obtained along 10 executions of our algorithm. Below, and for comparison, the values corresponding to the original filtered volume are shown.

TABLE 3 Alpha complex parameters of DnaBC

No. of Pseudo-Atoms	Alpha	No. of Tetrahedra	No. of Out. Triangles	Disk Space (Bytes)
200	4.79	319	382	3152
600	2.57	1903	1029	17024
1000	2.03	3820	1477	33560
1700	1.47	7691	1943	56860

Representative geometric and computational parameters of the series of alpha complexes obtained from the cryoelectron microscopy reconstruction of DnaBC shown in Fig. 7.

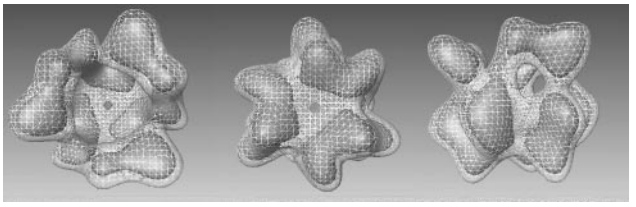

FIGURE 6 The DnaBC complex. Isosurface representation of DnaBC enclosing 100% of its molecular mass.

A_a^i. As it can be readily observed from the statistical data shown in Table 1 and 2, the alpha shape representation is quite stable with respect to preservation of shape and geometry information. In particular, the top row of Table 1 shows that the difference between the values of the histogram-based shape features calculated on a different realization of the alpha complex representation of Gal-6 is so small that they consistently have an effect only on the second decimal digit in a scale between 0 and 1 (the robustness is demonstrated by the small variance in the 10^{-5} range). Also, the similarity index with respect to the original object (Table 1, bottom row) remains virtually the same, with changes appearing only in the sixth decimal digit (in a scale from 0 to 1).

Similar results are shown in Table 2 for single-valued shape features. Indeed, the values of the features "size of the bounding box," "cross-correlation coefficient," and "eigenvalues of the inertia matrix" are very similar, with changes at most on the sixth decimal digit. Additionally, these values are consistently very close (within the second decimal digit at most) to those derived from the original low-resolution dataset.

Generation of the alpha complex representation from cryoelectron microscopy data

The second dataset used in this study corresponds to an experimental low-resolution map of DnaB and its loading partner DnaC obtained by cryoelectron microscopy and image processing techniques. DnaB is the main replicative helicase of *Escherichia coli*. In general, replicative helicases are motor proteins that unwind DNA at replication forks. Strand separation in DNA duplexes is a key step in many cellular events. Recently, the 3D low-resolution structure of *E. coli* DnaB hexamer in complex with its loading partner, DnaC, was obtained at 26 Å (Bárcena et al., 2001) (Fig. 6). The structure presents a global toroidal shape, with a central inner channel that is closed—at the resolution of this study—at one of the apical faces of the volume. The maximum external diameter is 13.8 nm and the reconstructed height is 12.4 nm.

This experimental low-resolution map was taken as input to the alpha shape representation generation algorithm presented in above. A number of alpha complexes were generated with an increasing number of pseudo-atoms, and a series of shape features were calculated on them. A gallery with these alpha complexes (constructed from 200, 600, 1000, and 1700 pseudo-atoms) is shown in Fig. 7. The final selected alpha shape representation corresponds to the one shown in Fig. 7 *d*, as this is the smallest (in terms of the number of pseudo-atoms) representation for which the average of all shape similarity indices with respect to the original low-resolution map is above 95%.

Note that the quality of surface representation used in Figs. 6 and 7 is different. While they visually differ, the reader should not mistakenly infer that there exists any difference in the data representation. In the first case, Fig. 6 is generated with a smooth rendering, while the second is a direct representation of the alpha complexes (triangular facets) that allow us to visualize the underlying alpha shape (without the additional step of smoothing for visualization).

It is interesting to study the evolution of the series of alpha complexes shown in Fig. 7, with supporting data in Tables 3 and 4. Table 3 presents a number of geometrical and computational parameters of the alpha complexes. It can be appreciated how the value of alpha decreases as the number of pseudo-atoms increases, which is a logical process considering that the alpha value was taken as the largest of the quantization errors in the generation of the pseudo-atoms. However, the complexity of the resulting alpha complexes increases (number of tetrahedra and triangles) as well as the total number of bytes needed to store the alpha complex. Regardless of this increase, the resulting representation is more compact than the original dataset because the final alpha complex requires about one-fourth of the space with respect to the original volume (a 64-cube coded at one byte per pixel).

As shown in Table 4, the general trend of all shape similarity measures presents a pattern of rapid increment followed by a slower phase of tuning. This fact is in accordance with the results previously obtained for Gal-6.

Automatic analysis of cavities, channels, and voids

Now, we focus on some advantages of handling the alpha complex instead of the original dataset. In particular, we

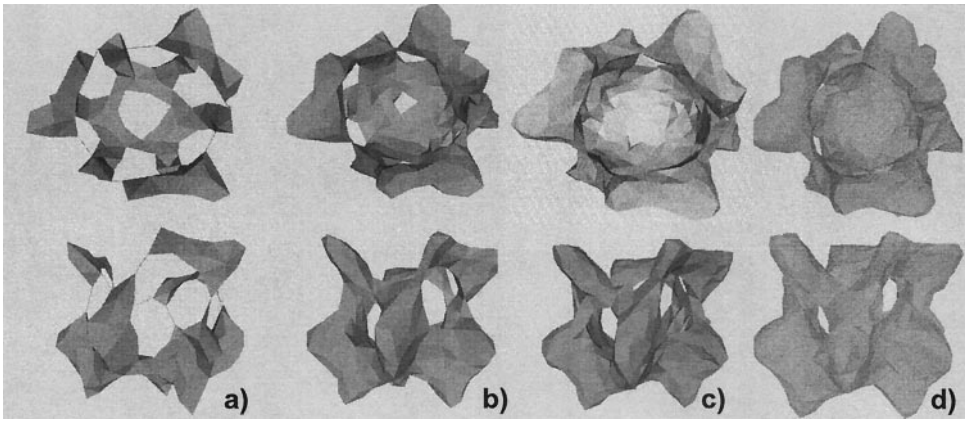

FIGURE 7 Evolution of the alpha-complex shape. The algorithm increases the cardinality of the point set to improve the fidelity of the corresponding alpha complex with respect to the original 3D map. Front and side views of alpha complexes with 200 pseudo-atoms (a), 600 (*b*), 1000 (*c*), and 1700 (*d*) are shown. The alpha complex (*d*) approximates the original model with an overall accuracy of 95% (from averaging all the shape feature scores), which satisfies our criterion of convergence.

provide the user with fully automatic extraction and analysis of cavities, channels, and voids that exist in the surface of the macromolecule.

Previously in this section it was noted that Gal-6 has an inner channel that could be involved in interaction with nucleic acids. In general, a desirable automatic analysis capability would yield means to detect and quantify channels of any number of mouths. In fact, this is a quite direct application of the alpha shapes theory (Edelsbrunner et al., 1995; Liang et al., 1998). Fig. 8, *top row*, presents a gallery of views of Gal-6 in which those tetrahedra that define the passing channel have been selected. Therefore, the particular shape features of the channel can now be studied in detail. Analogously, in Bárcena et al. (2001) it is described that the channel of DnaBC was closed at that resolution. Thus, a large cavity along the symmetry axes of the specimen was formed. Fig. 8, *bottom row*, shows a set of views of the alpha complex of DnaBC where solid tetrahedra fill the inner cavity. Now, it is straightforward to automatically analyze the local properties of this cavity (for instance, its total measured volume is 15,329 $Å^3$).

In a database context, the alpha shape representation allows us to easily pose complex geometric queries such as "Retrieve those specimens that present a passing channel with a diameter compatible with dsDNA."

The examples shown in Fig. 8 are not intended to be an exhaustive presentation of the potential applications opened by the alpha complex volume representation model proposed in this work. Instead, they are aimed at clearly illustrating some of their possibilities. They will be further explored and applied in future work.

CONCLUSIONS AND FUTURE WORK

In this work we propose a new way to represent low-resolution density maps that encodes, in a simple and compact manner, key information on their density, topology, and geometric characteristics. The proposed data representation allows for fast and accurate computations of macromolecular structural features while it significantly improves the required space that is needed to store each single dataset. It represents the first step toward modeling, storing, and efficiently querying volumetric images in a database infrastructure.

The approach consists of a two-step algorithm. First, a newly developed vector quantization technique termed kernel c-means is applied to select a point set (pseudo-atoms) within the macromolecule that best approximates the probability density function (pdf) of the density values of the original density map. The second step of the method is devoted to define connectivity relationships among points of the point set. To this end, the alpha shapes theory and public available software were used. It has been demon-

TABLE 4 Shape similarity measures calculated from different alpha shapes representations of DnaBC shown in Fig. 7

No. of ps-a	Iss (%)	Shell (%)	Sector (%)	Norm. *x* (%)	Norm. *y* (%)	Size (vox) (*z*, *y*, *x*)	Cross-Corr.	λ_0	λ_1	λ_2
Orig.	–	–	–	–	–	28,34,35	–	0.9489	0.9501	1
200	87.34	39.96	62.18	83.38	79.14	25,33,35	0.671	0.8890	0.8971	1
600	93.72	75.23	90.39	87.37	83.84	26,34,35	0.868	0.9427	0.9449	1
1000	96.39	81.03	95.61	90.19	87.48	26,34,35	0.915	0.9422	0.9488	1
1700	97.98	89.38	97.93	92.28	88.77	28,34,35	0.943	0.9450	0.9496	1

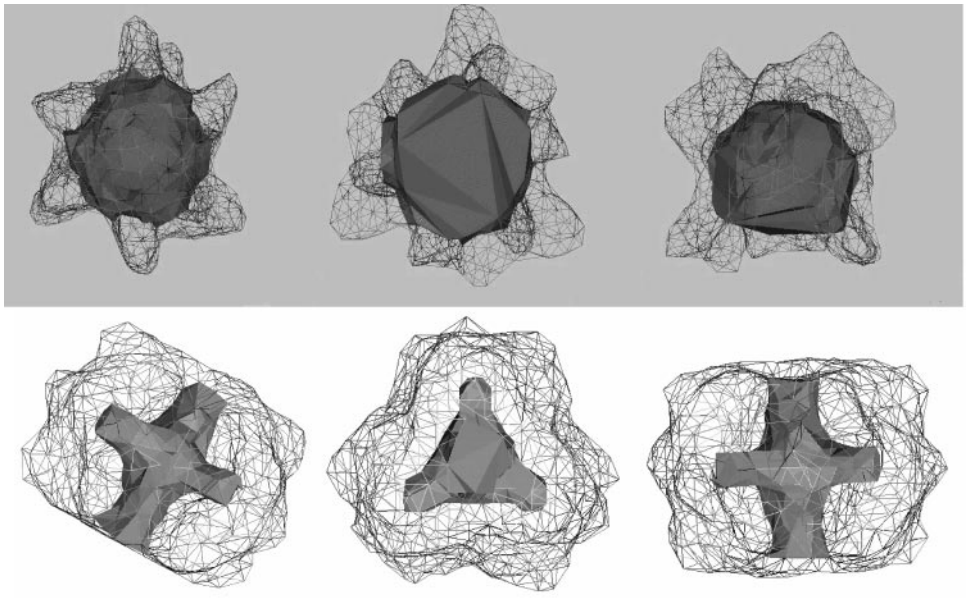

FIGURE 8 Automatic segmentation of passing channels and cavities of EM datasets. *Top row*: Views from different angles of the passing channel of Gal-6. It is opened to the exterior at two opposite sides of the molecular surface. *Bottom row*: Three different views of the open cavity of the DnaBC complex displayed with the alpha shapes suite of programs (http://alpha.geomagic.com/alpha/). Its measured volume is 15,329 $Å^3$.

strated that the alpha complex obtained using a value for alpha directly associated to the quantization error of the pdf estimation using kernel c-means closely approximates the shape and topology of the original molecule.

As a consequence, the resulting data structure is especially suitable for posing complex queries over large datasets of medium-resolution structures involving density, geometry, and topology characteristics. Queries such as "Retrieve those specimens that present a passing channel with a diameter compatible with dsDNA" can now be posed in an accurate and efficient manner. Other type of queries about shape similarity or about accessibility and measures of cavities are also possible.

However, it would be desirable to further refine the current data representation through iterative rounds of practical application and improvements. We realize the need to incorporate local shape features tailored to the biological meaning and role of channels and cavities. Also, the alpha shape software needs to be expanded to deal with shallow cavities and protrusions, as they are entities of biological importance. Research along these two lines is currently underway.

We thank Monica Chagoyen, Natalia Jimenez-Lozano, and Carlos Oscar Sanchez for fruitful discussions and helpful comments concerning the present work. We thank the reviewers of this article for the detailed revision and useful comments.

This work was supported in part by the Comision Interministerial de Ciencia y Tecnología through grants CICYT (P.N: Biotec., BIO 98-0761) and by the European Union through grants QLRI-CT-2000-31237 and QLRT-2001-00015. This work has been partly funded by the project TEMBLOR (The European Molecular Biology Linked Original Resources) through grant EU (QLRI-CT-2001-00015). The collaboration with the San Diego SuperComputer Center was supported by Grant HNCCT-99109 from the U.S.-Spain Science and Technology Program 1999, Ministerio de Asuntos Exteriores. P. A. De-A. is the recipient of a fellowship from Comunidad de Madrid (CAM).

REFERENCES

Akkiraju, N., H. Edelsbrunner, P. Fu, and J. Qian. 1996. Viewing geometric protein structures from inside a CAVE. *IEEE Comput. Graphics Appl.* 16:58–61.

Ankerst, M., G. Kastenmuller, H. P. Kriegel, and T. Seidl. 1999. Nearest-neighbor classification in 3D protein databases. *Proc. 7th Int. Conf. Intell. Syst. Mol. Biol.* (ISMB '99) Heidelberg, Germany. 34–43.

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

Bárcena, M., T. Ruiz, L. E. Donate, S. Brown, N. Dixon, M. Radermacher, and J. M. Carazo. 2001. The DnaB-DnaC complex: a structure based on dimers assembled around an occluded channel. *EMBO J.* 20:1462–1468.

Baumeister, W., and A. C. Steven. 2000. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.* 25: 624–631.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. Shindyalov, and P. E. Bourne. 2000. *Nucleic Acids Res.* 28:235–242.

Bezdek, J. C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York.

Bohm, J., A. S. Frangakis, R. Hegerl, S. Nickell, D. Typke, and W. Baumeister. 2000. Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl. Acad. Sci. U.S.A.* 97:14245–14250.

Canny, J. 1986. A computational approach to edge detection. *IEEE Trans. Pat. Anal. Machine Intell.* 8:679–698.

Connolly, M. L. 1983a. Analytical molecular surface calculation. *J. Appl. Crystallogr.* 16:548–558.

Connolly, M. L. 1983b. Solvent-accessible surfaces of proteins and nucleic acids. *Science*. 221:709–713.

Connolly M. L., T. O'Donnell, and S. Warde. 1996. Special issue on molecular surfaces. *Network Sci.* 2:4.

Delfinado, C. J. A., and H. Edelsbrunner. 1995. An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere, Grid gener-

ation, finite elements, and geometric design. *Comput. Aided Geom. Design.* 12:771–784.

Deriche, R. 1987. Using Canny's criteria to derive a recursively implemented optimal edge detector. *Image and Vision Computing.* 1:167–187.

Edelsbrunner, H., and E. P. Mucke. 1994. Three-dimensional alpha shapes. *ACM Trans. Graphics.* 13:43–72.

Edelsbrunner, H., M. A. Facello, P. Fu, and J. Liang. 1995. Measuring proteins and voids in proteins. *Proc. 28th Annu. Hawaii Internat. Conf. System Sciences, V.* 256–264.

Edelsbrunner, H., M. A. Facello, and J. Liang. 1998b. On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* 88:83–102.

Edelsbrunner, H., J. Liang, and C. Woodward. 1998a. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* 7:1884–1897.

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

Gersho, A., and R. M. Gray. 1992. Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston.

Grimes, J. M., S. D. Fuller, and D. I. Stuart. 1999. Complementing crystallography: the role of cryo-electron microscopy in structural biology. *Acta Crystallogr. D. Biol. Crystallogr.* 10:1742–1749.

Holm, L., and C. Sander. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123–138.

Joshua-Tor, L., E. H. Xu, S. A. Johnston, and D. C. Reeds. 1995. Crystal structure of a conserved protease that binds DNA: the bleomycin hydrolase, Gal-6. *Science.* 269:945–950.

Kalko, S. G., M. Chagoyen, N. Jimenez-Lozano, N. Verdaguer, I. Fita, and J. M. Carazo. 2000. The need for a shared database infrastructure: combining x-ray crystallography and electron microscopy. *Eur. Biophys. J.* 29:457–462.

Keller, P. A., K. Henrick, P. McNeil, S. Moodi, and G. J. Barton. 1998. Deposition of macromolecular structures. *Acta Crystallogr. D. Biol. Crystallogr.* 54:1105–1108.

Liang, J., H. Edelsbrunner, and C. Woodward. 1998. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* 7:1884–1897.

Lohmann, G. 1998. Volumetric Image Analysis. Wiley-Teubner.

Morris, G. M., A. J. Olson, and D. S. Goodsell. 2000. Protein-ligand docking. *In* Evolutionary Algorithms in Molecular Design, D. E. Clark, editor Wiley-VCH, Weinheim, Germany. 31–48.

Murata, K., K. Mitsuoka, T. Hirai, P. Agre, J. B. Heymann, A. Engel, and Y. Fujiyosi. 2000. Structural determinants of water permeation through aquaporin-1. *Nature.* 407:599–605.

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

Norel, R., D. Petrey, H. J. Wolfson, and R. Nussinov. 1999. Examination of shape complementarity in docking of unbound proteins. *Proteins.* 36:307–317.

Paquet, E., and M. Rioux. 1998. Content-based access of VRML libraries. IAPR International Workshop on Multimedia Information Analysis and Retrieval. Lecture Notes in Computer Sciences. Springer 20–32.

Parzen, E. 1962. On the estimation of a probability density function and the mode. *Ann. Math. Stat.* 33:1065–1076.

Pascual-Montano, A., L. E. Donatec, M. Valle, M. Bárcena, R. D. Pascual-Marqui, and J. M. Carazo. 2001. A novel neural network technique for analysis and classification of EM single particle images. *J. Struct. Biol.* 133:233–245.

Peters, K. P., J. Franck, and C. Frommel. 1996. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* 256:201–213.

Shindialov, I. N., and P. E. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11:739–747.

Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Sussman, J. L., D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola. 1998. Protein data bank (PDB): database of 3D structural information of biological macromolecules. *Acta Crystallogr. D.* 54: 1078–1084.

Volkmann, N., and D. Hanein. 1999. Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J. Struct. Biol.* 125:176–184.

Westhead, D. R., T. W. Slidel, T. P. Flores, and J. M. Thornton. 1999. Protein structural topology: automated analysis, diagrammatic representation and database searching. *Protein Sci.* 8:897–904.

Wriggers, W., and S. Birmanns. 2001. Using Situs for flexible and rigid-body fitting of multiresolution single-molecule Data. *J. Struct. Biol.* 133:193–202.

Wriggers, W., R. A. Milligan, and A. McCammon. 1999. Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* 125:185–195.

Wriggers, W., R. A. Milligan, K. Schulten, and A. McCammon. 1998. Self-organizing neural networks bridge the biomolecular resolution gap. *J. Mol. Biol.* 184:1247–1254.