Journal of Computer-Aided Molecular Design, 10 (1996) 55–66
ESCOM

J-CAMD 323

Similarity and complementarity of molecular shapes: Applicability of a topological analysis approach*

Laurence Leherte**, Thibaud Latour and Daniel P. Vercauteren

Laboratory of Computational Chemical Physics, Institute for Studies in Interface Sciences, University of Namur, Rue de Bruxelles 61, B-5000 Namur, Belgium

Received 15 October 1995
Accepted 18 November 1995

Keywords: Critical points; Electron density; Proteins; Cyclodextrins

Summary

Developments based on a topological analysis approach of electron density maps are presented and applied to two different fields: the interpretation of electron density maps of proteins and the description of shape complementarity between a cyclodextrin host and a guest molecule. A global representation of the electron density distribution, through the location, identification and linkage of its critical points (points where the gradient of the density vanishes, i.e., peaks and passes), is generated using the program ORCRIT. On one hand, the interpretation of protein electron density maps is based on similarity evaluations between graphs of critical points and known structures. So far, the method has been applied to 3 Å resolution maps for the recognition of secondary structure motifs using a procedure relevant to expert systems in artificial intelligence. Satisfying matches between critical point graphs and their corresponding protein structure depict the ability of the topological analysis to catch the essential secondary structural features in electron density maps. On the other hand, mapping the accessible volume of a host molecule is achieved by representing the peaks as ellipsoids with axes related to local curvature of the electron density function. Related energies of the interacting species can also be estimated. A qualitative comparison is made between the results generated by the topological analysis and energy values obtained by conventional molecular mechanics calculations. A positive comparison and a close complementarity between cyclodextrin and ligands shows that the topological analysis method gives a good representation of the electron density function.

Introduction

Several years ago, Johnson developed a topological analysis approach dedicated to the representation of a protein structure in an electron density map (EDM) in terms of its critical points and their linkage. This critical point approach was first implemented in the computer program ORCRIT [1], which was developed as part of the Crysalis project [2]. The purpose of that project was to build an expert system for the automated interpretation of experimental protein EDMs in the 2.0–2.5 Å resolution range. More recently, we have applied the same program to maps at medium resolution, i.e., 3 Å, to the interpretation of protein structures in the framework of the Molecular Scene Analysis project [3], and to the evaluation of steric complementarity between netropsin and a B-DNA dodecamer structure [4]. We have assessed its usefulness as a tool for segmenting the maps and interpreting the segmented fragments. This study has led to extensions of the original ORCRIT implementation. In the present work, we wish to use the local information associated with the critical point representations to assess the capabilities of this particular topological approach in structure and volume representations.

The first application reported in this paper is relevant to the identification of protein secondary structure motifs in EDMs. The calculation of similarity degrees between 3D properties of molecules usually calls for knowledge of

*This paper is based on a presentation given at the 14th Molecular Graphics and Modelling Society Conference, held in Cairns, Australia, August 27–September 1, 1995.
**To whom correspondence should be addressed.

the molecular surfaces [5] and/or the electron density functions calculated through the use of quantum-mechanical calculations, and related properties such as the molecular electrostatic potential and the electric field [6]. In such studies, the comparison between two species is usually based on a 3D property, the same for both partners. In the present study, we wish to compare two different representations, i.e., the electron density function associated with a protein structure and its molecular backbone. Electron density functions are generated by a method simulating X-ray diffraction experiments, which involves a drastic gain of computer time but less accuracy. ORCRIT is not the only skeletonization procedure considered in the crystallographic community. Other graph analysis methods that rely on visualization tools are currently used [7]. However, ORCRIT has the advantage to classify the vertices of the graph it generates by using a mathematical operator, the Laplacian operator ∇^2, which acts as a filter to naturally convert a continuous representation, i.e., a 3D electron density function, to a discrete, rather symbolic representation. For example, the sign of $\nabla^2\rho$ may be considered to classify critical points and to segment a 3D function into separate regions. This idea was also developed by Marr in his work on vision [8]. In the present work, we study the nature of the critical points in protein EDMs and the connectivity between these critical points in order to determine its representativity of the protein structure.

The second application presented in this paper deals with the analysis of the local properties of critical points in order to determine whether they involve sufficient information to represent the actual spatial distribution of the electron density. There are many levels of representation of a molecular shape, most of them using graphics interfaces for visualization. We are particularly interested in shape comparisons without any graphics assistance, i.e., we need a way to easily and quantitatively characterize 3D surfaces for quick comparisons. Descriptions of molecular shapes are often based on the use of hard-sphere representations, since modelling of molecular recognition processes usually involves a combination of purely steric interactions with intermolecular potentials resulting from hydrophobic and electrostatic effects, as well as hydrogen bonding. For example, the solvent-accessible portion of a molecule represented as van der Waals (vdW) spheres can be determined by considering the locus of the centre of a spherical probe of appropriate size sweeping the surface [9]. Derived methods have also been proposed. Del Carpio et al. [10] have represented atoms as sets of points defining an icosahedral body embedded in each vdW sphere for the identification of cavities of proteins hosting a ligand molecule. In order to determine the optimum interaction between two molecules composed of hard spheres, Santavy and Kypr [11] defined the maximal interface area by fitting projections of the molecular surfaces onto grids and minimising the surface–surface distances. Arteca and Mezey [12] additionally considered the interrelation between vdW spheres and segmented the molecular surfaces into sections of spherical face, while Lin et al. [13] characterized solvent-accessible (Connolly) surfaces. In the latter work, a critical point has a particular meaning, i.e., it is the projection of the gravity centre of a Connolly section onto the underlying molecular surface. A method based on spherical harmonics was also developed by Leicester et al. [14] and applied to the description of Connolly surfaces, but the authors also suggested that the method is applicable to other surfaces, e.g., potential energy contours. The evaluation of shape complementarity between a compound and a receptor can for example be achieved using methods as developed in the program DOCK [15]. Such an approach was recently applied to the search of DNA-binding compounds [16] in the Cambridge Structural Database (CSD) [17]. However, here we wish to consider shape deformation due to anisotropy of the subunits within a molecule (atoms, functional groups, side chains, etc.) arising from chemical bonding and thermal motion. We have thus focussed on electron density analysis and interaction potential evaluation.

Materials and Methods

In this section, we describe the topological approach that is presently used, i.e., the determination of critical points and the establishment of their connectivity. The two derived methods that were briefly presented in the Introduction will be described in more detail in the next two sections.

Topological analysis approach

The charge distribution of a molecular system is characterized by an accumulation of electron charge at the nucleus positions. This is clearly observed in electron density distributions generated by quantum-mechanics methods [18,19]. It is also put in practice in crystal structure determination processes, where a model of the crystalline structure is built from the high-density regions found in experimental electron density maps (EDMs). To achieve a more detailed analysis of calculated charge distributions, a topological approach was proposed by Bader [20], which requires the computation of the gradients $\nabla\rho(\mathbf{r})$, the Hessians $\mathbf{H}(\mathbf{r})$, and the Laplacians $\nabla^2\rho(\mathbf{r})$ of the electron density. The eigenvalues of $\mathbf{H}$ provide information on the local curvature and the negative or positive Laplacians give details about the local concentration or depletion, respectively, of electron density.

The Hessian matrix of a continuous 3D function, such as the electron density distribution ρ, is precisely defined at locations $\mathbf{r}$ of its critical points, and is built from the second derivatives:

$$\mathbf{H}(\mathbf{r}) = \begin{pmatrix} \frac{\partial^2}{\partial x^2}\rho & \frac{\partial^2}{\partial x \partial y}\rho & \frac{\partial^2}{\partial x \partial z}\rho \\ \frac{\partial^2}{\partial y \partial x}\rho & \frac{\partial^2}{\partial y^2}\rho & \frac{\partial^2}{\partial y \partial z}\rho \\ \frac{\partial^2}{\partial z \partial x}\rho & \frac{\partial^2}{\partial z \partial y}\rho & \frac{\partial^2}{\partial z^2}\rho \end{pmatrix} \quad (1)$$

This real and symmetric matrix can be put in a diagonal form. This corresponds to finding a rotation of the original coordinate system that aligns the new coordinate axes with the principal axes of the critical points. The diagonalized form is:

$$\mathbf{H}'(\mathbf{r}) = \begin{pmatrix} \frac{\partial^2}{\partial x'^2}\rho & 0 & 0 \\ 0 & \frac{\partial^2}{\partial y'^2}\rho & 0 \\ 0 & 0 & \frac{\partial^2}{\partial z'^2}\rho \end{pmatrix} \quad (2)$$

where each element of **H'** is the curvature of $\rho(\mathbf{r})$ with respect to a principal axis. The three non-zero diagonal elements are called the eigenvalues and the Laplacian is the trace of **H'**. The rank of **H'** is the number of non-zero eigenvalues, and the signature s is the algebraic sum of the signs of the eigenvalues. If the rank of **H'** is 3, then four cases are met: $s=-3$ corresponds to a local maximum or *peak*, i.e., the electron density function adopts maximum values along each of the three principal directions x', y' and z'; $s=-1$ corresponds to a saddle point or *pass*, where two of the eigenvalues are negative; $s=+1$ corresponds to a saddle point or *pale*, characterized by only one negative eigenvalue, and $s=+3$ corresponds to a local minimum or *pit*, i.e., the electron density function adopts minimum values along each of the three principal directions.

The points where the gradient of the electron density vanishes are known as the critical points of the electron density function. At atomic resolution, peaks and passes are normally associated with the presence of atoms and chemical bonds, respectively, while pales and pits occur as a result of the geometrical arrangement of the atoms and the corresponding networks of bonds. For example, pales and pits are found in the interior of rings and cages, respectively.

In the case of crystals, the critical point network is defined not only by the molecular structure but also by the lattice periodicity and the space group symmetry [21]. Due to periodic boundary conditions, a unit cell can be considered as a 3D torus, each pair of opposite faces being connected. This means that the motif of critical points is not isolated, but interacts with its periodic images and, therefore, the number of critical points is constrained by the following relationship:

$$n - b + r - c = 0 \quad (3)$$

where n, b, r and c stand for the numbers of peaks, passes, pales and pits, respectively. Thus, within a crystal structure, two sub-networks are intercalated. One of them is composed of peaks and passes, the other of pits and pales.

When the electron density distribution to be analyzed is precise enough, for example, when it is computed at atomic resolution by using quantum chemistry methods, several atomic properties can be derived by integrating the electron density distribution within the so-called zero-flux surfaces [20]. These surfaces are defined by the following relationship:

$$\nabla\rho(\mathbf{r}) \cdot \mathbf{n} = 0 \quad (4)$$

where **n** is a normal vector to the tangent plane defined at **r**. An analytical description of these surfaces has further been given by Popelier [22].

As specified previously, surfaces of constant electron density values also yield information about the shape of the molecular envelope. For example, in Mezey's approach [5], Hessian matrices are built in a local system of coordinates spanning a plane tangent to the constant electron density surface. The signs of the eigenvalues are determined at each point on the surface. A set of adjacent points for which two eigenvalues are positive forms a *convex* domain. If only one eigenvalue is negative, the domain is *saddle-type*, and if none of the eigenvalues is negative, the points define a *concave* domain. Mezey's method is very useful for the analysis of electron density distributions, as long as density values remain precise, even away from the high-density centres. This criterion, however, is not satisfied by our reconstructed EDMs. Finally, the critical point approach can be applied not only to the analysis of the electron density distribution but also to other functions, such as the molecular electrostatic potential, as in the studies proposed by Gadre et al. on small inorganic molecules [23].

The connection pattern linking critical points can be determined by following the density gradient vector $\nabla\rho(\mathbf{r})$ [20]. In ORCRIT, an alternative approach is followed. The characteristics of nearby critical points – types, distances, density heights, eigenvector projections – are used in order to derive a weight function, which is then used to establish possible connections. In particular, such a function can be used in a graph theoretical approach for representing the full network of critical points. A graph is a representation of a finite set of points (or vertices) and their connectivity, i.e., the corresponding set of edges. In particular, a connected graph consists of a set of vertices in which any pair may be connected by a path formed by one or several successive edges in the graph. When a graph includes closed paths (circuits), it is possible to remove one edge from each circuit, resulting in a

graph for which all points still remain connected. This graph, from which ring-closure connections are removed, is a spanning tree. The tree is called a minimal spanning tree if, among all of the possible spanning trees that can be formed, the sum of the weights associated with its edges has a minimal value. For this reason, the weight of a connection between two critical points, w_{ij}, as calculated by ORCRIT, is inversely proportional to its occurrence probability. A set of spanning trees is described as a forest.

ORCRIT analyses the set of critical points so as to generate a forest of minimal spanning trees. The weight w_{ij} of any pair of critical points is computed according to the following equation:

$$w_{ij} = \Delta r \, R \, F \, \Delta s \qquad \text{where } R = \frac{1 + (|\Delta\rho|/(\rho_i + \rho_j))}{\sqrt{\rho_i + \rho_j}} \qquad (5)$$

Here, Δr is the distance between the two critical points, Δs is equal to 1.1, 1.0, or 5.0, depending on whether the difference in the number of negative eigenvalues of the two critical points is equal to 0, 1, or greater than 1, respectively, and F is equal to $(3 - F_i - F_j)$, where F_i denotes the normalized projection of the eigenvector of point i along the vector Δr connecting the two critical points. The Δs term prevents connections from being traced between peaks and pales or pits. It also permits connections between two peaks even when the pass in-between them has not been detected. The term R favours connections between high-density critical points of similar heights. Also, the weight w_{ij} is set equal to Δr when the distance Δr is smaller than 0.25 Å. This is done in order to force a connection between symmetry-equivalent critical points.

Materials

Ideal maps at 3 Å resolution were considered for the following reasons. First, protein diffraction data are often noisy due to phase determination errors, and out of atomic resolution range. The Molecular Scene Analysis project thus considers the interpretation of EDMs as a hierarchical problem, where finer structural details are iteratively inferred from less well resolved maps [24]. A second reason lies in the fact that the interpretation method should be highly automated, i.e., the representations should involve a reduced number of variables as compared to the number of all atom coordinates. The use of a hierarchical approach to reduce the cost of molecular simulations was also proposed by Butzlaff et al. [25], by approximating energy functions and force field constants for DNA models decomposed into subunits. Hierarchical representations are particularly suited to biological macromolecules such as proteins and DNA, because of their regular structure.

All maps analyzed in this paper were reconstructed from coordinates retrieved from the Brookhaven Protein Data Bank (PDB) [26] or the CSD [17], and were submitted as input to XTAL, a crystallographic resolution package [27].

High-density peaks and passes were the only critical points considered in this study. Low-density critical points are less significant, because the electron density distribution is modulated by experimental noise and/or errors due to the fast Fourier transform process. In addition, the analysis levels some low-density points (those with negative values) to zero. High-density peaks and passes were considered by imposing a cutoff value below which the critical point search procedure is not applied.

Secondary structure recognition in protein electron density maps (EDMs)

This work was initiated by Glasgow et al. [24] to provide a tool for the automated interpretation of experimental EDMs of protein structures. This is actually a complex problem, requiring both expert knowledge and sophisticated computational techniques. To address both of these requirements, the Molecular Scene Analysis project turned to artificial intelligence techniques [24]. A preliminary but important step in automated map interpretation concerns the implementation of approaches that can be used for the recognition and classification of structural motifs, e.g., secondary structures, residues, etc., at varying levels of resolution. An adequate and easily implementable representation of the information is thus required to provide a classification scheme of the various structural motifs, and to construct templates or identify rules that

TABLE 1
DENSITY CUTOFF, MAXIMUM ALLOWED WEIGHT AND LENGTH OF CONNECTIONS, AND GRID INTERVAL USED FOR THE CONSTRUCTION OF CRITICAL POINT GRAPHS[a]

PDB structure	No. of residues	Density cutoff ($e^-/Å^3$)	Weight	Length (Å)	Grid interval (Å) along		
					a	b	c
4PTI	58	1.0	4	4	0.5	0.5	0.5
1RNT	104	1.0	4	4	0.5	0.5	0.5
1BP2	123	1.0	4	4	0.8	0.6	0.6
3APP	323	1.0	4	4	0.9	0.9	0.8

[a] The data were used as ORCRIT input parameters for the analysis of protein electron density maps of 3 Å resolution.

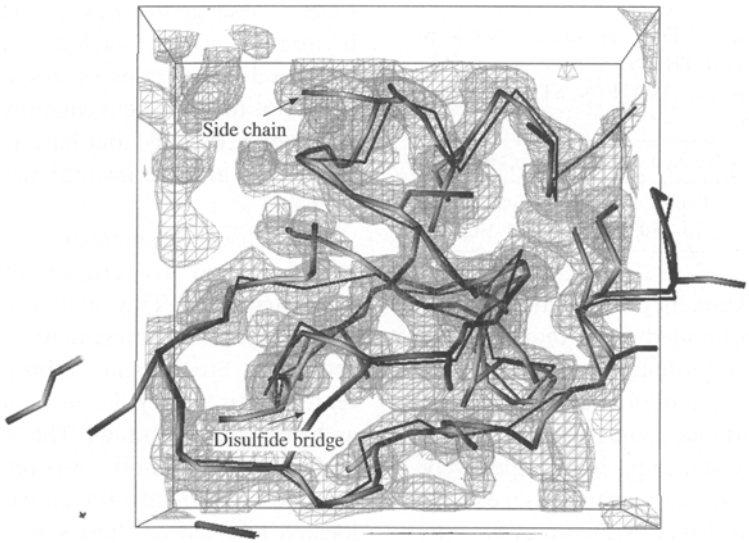

Fig. 1. Superimposition of the critical point tree (peaks: thick dark segments; passes: thick grey segments) and $C^{\alpha}CO$ centres of mass (thin black lines) of protein structure 4PTI as obtained using ORCRIT on a reconstructed density map of 3 Å resolution (the grey net contour level is 0.95 $e^{-}/Å^{3}$). The picture was drawn using IBM Visualization DataExplorer, v. 1.2 (IBM Corporation, Yorktown Heights, NY, 1992).

can be used for the recognition and interpretation of EDMs. In this section, we describe how critical point representations, obtained using ORCRIT, fulfill these two requirements. The maps were reconstructed for four structures stored in the PDB [26]. Their identification codes are: 4PTI, 1RNT, 1BP2 and 3APP. Input parameters needed to run ORCRIT are specified in Table 1.

As previously shown, peaks correspond to amino acid residues along the main chain of the protein and passes to the connectivity determined by the primary structure of the protein [3]. As illustrated in Fig. 1, the topological approach produces a skeleton of the protein structure as a sequence of alternating peaks and passes, where each peak is associated with one residue of the protein. For larger residues, side chains are also included in the tree. The results obtained from the analysis of calculated EDMs led to the following four observations. (i) The main branch of the spanning tree traces out the backbone of the protein molecule. Peaks are located close to the C^{α}-C=O centres of charge for the residue, at an average distance of 0.55 ± 0.32 Å. (ii) Each peak of the main branch of the tree is associated with a single residue of the primary sequence for the protein. (iii) Side chains are often observable, particularly for the larger residues. These chains are represented as side branches, which link to the main branch of the spanning tree. (iv) Disulfide bridges are mostly discerned through the highest density values of their peaks. Peaks associated with cysteine and methionine side chains always present the highest magnitudes.

The result of applying the ORCRIT program is thus a partitioning of the EDM into two main regions: the protein region represented by a chain of connected critical points, and a solvent region that is characterized by low density values and non-connected critical points. The chain of critical points itself denotes a segmentation of the protein into meaningful parts, i.e., amino acid residues and peptidic bonds. Thus we have considered an identification procedure of secondary structure motifs in a medium EDM, based on these observations and conclusions. It is presented below.

Similarity analysis

A statistical analysis of the conformation of critical point sequences in terms of geometrical parameters of four peak segments (p_i - p_{i+1} - p_{i+2} - p_{i+3}) showed that the most useful parameters for the identification of helices and β-strands were the torsion angles, the distances between peaks p_i and p_{i+3}, and the bond angle values [3]. Peaks only are involved in the following procedure if they have been identified as meaningful features.

The recognition of secondary structure features from medium-resolution EDMs is based on pattern matching of the critical point network onto idealized secondary structure motifs. These idealized motifs, or templates, can be limited in size due to the regular and repetitive shape of helices and β-strands. α-Helices occurring in natural proteins have 3.6 residues per turn, whereas the less frequent 3_{10}-helices are characterized by 3 residues per turn. The templates we used are actually chains composed of four points, each point representing a C^{α}-C=O centre of charge.

A previous approach [28] has treated these geometrical constraints in an intuitive way, i.e., they were either satisfied or not, depending upon whether the values fall within predetermined ranges. The procedure presented in this paper is less intuitive, since the measure of belief it associates with a given hypothesis depends on information statistically derived from the Brookhaven PDB [26].

TABLE 2
MOST COMMON GEOMETRIES FOR HELICES AND β-STRANDS AS OBTAINED FROM THE STATISTICAL ANALYSIS OF 63 NONHOMOLOGOUS PROTEIN STRUCTURES RETRIEVED FROM THE PDB

Template (model)	Bond angle (°)	Torsion angle (°)	$d(p_i - p_{i+3})$ (Å)
Helix	80–90	50–60	4.5–5.0
β-strand	120–130	|170–180|	10.0–10.5

First, a set of occurrence frequency distributions p(ssm, g) and p(not_ssm,g) was computed from the statistical analysis of 63 nonhomologous protein structures retrieved from the PDB [26]. ssm and g stand for secondary structure motif (helix or β-strand) and geometrical parameter (torsion angle, bond angle, distance $p_i - p_{i+3}$), respectively. These frequency distribution values were used as measures of belief MB(ssm,g) and disbelief MD(ssm,g) for the secondary structure assignment, given a particular geometry parameter g for each four-peak segment. Several methods for determining confidence values for secondary structure assessment were considered. We found a method similar to the one applied in the diagnosis system MYCIN, described in Ref. 29, effective for our applications. This method is classified among the 'ad hoc' approaches for dealing with uncertain knowledge [30]. Other statistical methods, such as the Dempster–Schafer theory, are applied to deal with uncertainty reasoning for 3D object recognition [31]. In the latter study, assigning belief to objects consisted in comparing attributes of their shape, relationships, and aspect to attributes of models. In our case, we do not have to deal with the objects (critical points) themselves but only with their spatial relationships. However, the 3D shape of each peak is not absent, as it is stored in terms of the three local density curvatures.

In our process, when the combination of two pieces of evidence ($g_1 \Lambda g_2$) was required, the calculation of the MBs and MDs was achieved using the following two formulas:

$$\mathrm{MB}(\mathrm{ssm}, g_1 \Lambda g_2) = \mathrm{MB}(\mathrm{ssm}, g_1) + \mathrm{MB}(\mathrm{ssm}, g_2)(1 - \mathrm{MB}(\mathrm{ssm}, g_1)) \quad (6)$$

$$\mathrm{MD}(\mathrm{ssm}, g_1 \Lambda g_2) = \mathrm{MD}(\mathrm{ssm}, g_1) + \mathrm{MD}(\mathrm{ssm}, g_2)(1 - \mathrm{MD}(\mathrm{ssm}, g_1)) \quad (7)$$

where MB is a measure of belief in the hypothesis ssm given the evidence g, while MD measures the extent to which evidence g supports the negation of the hypothesis ssm. Given these measures, an overall certainty factor, CF, in the range −1 to 1 was determined for each peak as the difference between the final belief and disbelief measures.

In order to check that MB tends to 1 when submitted to the recognition of ideal motifs, the recognition procedure was applied to templates with most common geometries as reported from the statistical analysis of the PDB. These geometries are reported in Table 2. The highest CF values reached were 1.00 and 0.97 for the helix and strand models, respectively. However, we decided not to limit the recognition evaluators to the calculation of MBs only, and have considered the MD values because, in an experimental map, data are expected to be fuzzier.

A result of the application of this method to the interpretation of an ideal critical point tree is shown in Fig. 2 for protein 4PTI. This small protein structure was selected for clarity of the presentation of the results. It clearly shows that broad bands of positive CF values are indeed representative of regular secondary structure motifs, such as helices and β-strands. The location of the actual secondary structure motifs was retrieved from the PDB file, i.e., one 3_{10}-helix, two β-strands, one turn and one α-helix located between residues 3–6, 16–23, 28–35, 23–27, and 47–56, respectively. Both kinds of helices, 3_{10} and α, are identified but not strictly differentiated, except that 3_{10}-helices give rise to lower CF values. This is due to deviations in the geometrical features of the 3_{10}-helices, which reduce the final MB values. The two β-strands occurring in protein structure 4PTI are clearly observed, and are separated by a turn that was identified using the hypothesis 'helix'. In this context, short helices and turns will be differentiated with difficulty. Differentiation should thus occur at a higher resolution level.

Evaluation of shape complementarity for host–guest systems

The method described in this section was applied to the study of confined systems, i.e., guest molecules interacting with a highly concave host surface. It is illustrated by an application to chiral flurbiprofen/β-cyclodextrin complexes. The host and guest chemical structures are

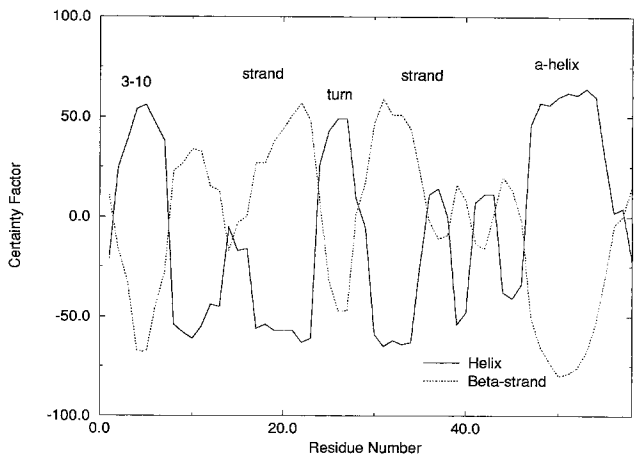

Fig. 2. Certainty factor values (scaled by 100) obtained from the application of the MYCIN algorithm to hypotheses 'helix' and 'strand' with the idealized critical point tree representation of the protein 4PTI (58 amino acids).

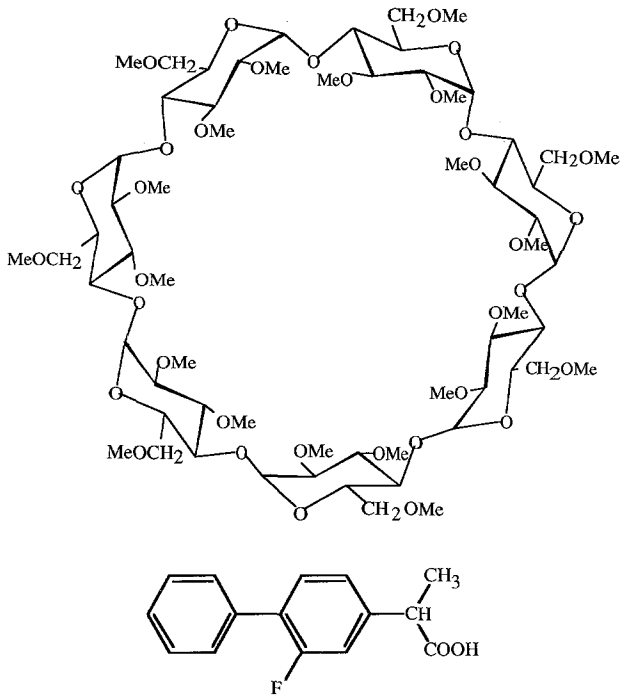

Fig. 3. Chemical structures of the permethylated β-cyclodextrin (top) and flurbiprofen (bottom) molecules.

depicted in Fig. 3. Crystal structures of the *R*- and *S*-complexes were retrieved from the CSD [17], codes COYXAP10 and COYXET20, respectively. The crystal structure of the *R*-complex includes one water molecule, whereas the *S*-complex is unhydrated. In both complexes, the flurbiprofen phenyl group is inserted into the cyclodextrin cavity while the carboxyl group, which is bonded to the asymmetric carbon atom, is located outside the cavity and interacts with another cyclodextrin molecule. EDMs of both complexes were reconstructed as described in the Materials section, and analyzed using ORCRIT [1] with the following parameters: density cutoff = 1.0 $e^-/Å^3$, maximum weight = 4, maximum length = 4 Å, and grid interval = 0.35 Å.

At the critical point locations, the three main curvatures of the electron density function are the eigenvalues of the Hessian matrix constructed from the second derivatives. It is assumed that this local information can be transferred to the space surrounding the critical point concerned; hence it is possible to evaluate (or reconstruct) the 3D function in the close neighbourhood of each point. Each maximum of the electron density function, i.e., each peak, is considered as the centre of expansion of a Gaussian function and such a mathematical expression is fitted in order to define a volume around each peak, taking into account its three characteristic eigenvalues:

$$\rho(r) = \rho(0)\exp(\alpha \mathbf{r}^T \mathbf{H} \mathbf{r}/\rho(0)) \quad (8)$$

Within the frame of ORCRIT, the parameter α is equal to 2.0. In order to evaluate the volume associated with a particular peak, the exponential term of the Gaussian function has been integrated over the space within the frame of the ellipsoid:

$$\int_V \exp\left(\frac{\alpha \mathbf{r}^T \mathbf{H}' \mathbf{r}}{\rho(0)}\right) \quad (9)$$

which leads to the definition of an ellipsoid characterized by three main axes r_x, r_y, and r_z:

$$V = \frac{\pi^{3/2}\rho(0)^{3/2}}{2^{3/2} h_x^{1/2} h_y^{1/2} h_z^{1/2}} = \frac{4\pi}{3} r_x r_y r_z \quad (10)$$

and hence provides a method of representing shape anisotropy of the critical points. This shape description is extended to a whole molecule by considering a set of ellipsoids, and a descriptor for this structure is defined in terms of interaction energy values. This 'signature potential' is constructed from the lengths obtained for the ellipsoid axes and is therefore orientation-dependent. The total interaction between the host and another ellipsoid is expressed within a pair-potential approximation, where the dispersive interaction between two ellipsoids i and j is proportional to their volume product:

$$U = \sum_{i<j} U_{ij} \quad (11)$$

$$U_{ij} = -A_{ij}/r_{ij}^6 + B_{ij}/r_{ij}^{12} \quad \text{where } A_{ij}>0 \text{ and } B_{ij}>0 \quad (12)$$

$$A_{ij} \approx V_i V_j \quad (13)$$

$$B_{ij} = A_{ij}(r_i + r_j)^6 \quad (14)$$

where r_{ij} is the separation distance between particles i and j, and r_i and r_j are their radii calculated along the inter-distance vector i–j. In such a formula, the equilibrium distance between i and j is considered to be given by $2^{1/6}(r_i+r_j)$. Thus, overlap between two individual particles is forbidden. The radius r_j is selected by the user or determined by the ellipsoid size. The idea of using a 'hypothetical' potential energy function in order to determine the optimal steric location, sterically speaking, of a guest molecule was also developed by Kuntz et al. [15] in the program DOCK. These authors simplified the overlap energy between two molecules to a contribution depending upon the vdW radii of the interacting atoms and their separation distance. However, considering a Lennard-Jones-type potential allows us to emphasize the effect of global curvature of the neighbourhood, e.g., a cavity leading to more attractive energies.

Fitting Gaussian functions to a higher resolution representation, i.e., to atoms, has been done later by Grant and Pickup [32] to overcome the limitations of hard-sphere representations of molecular shapes. From such

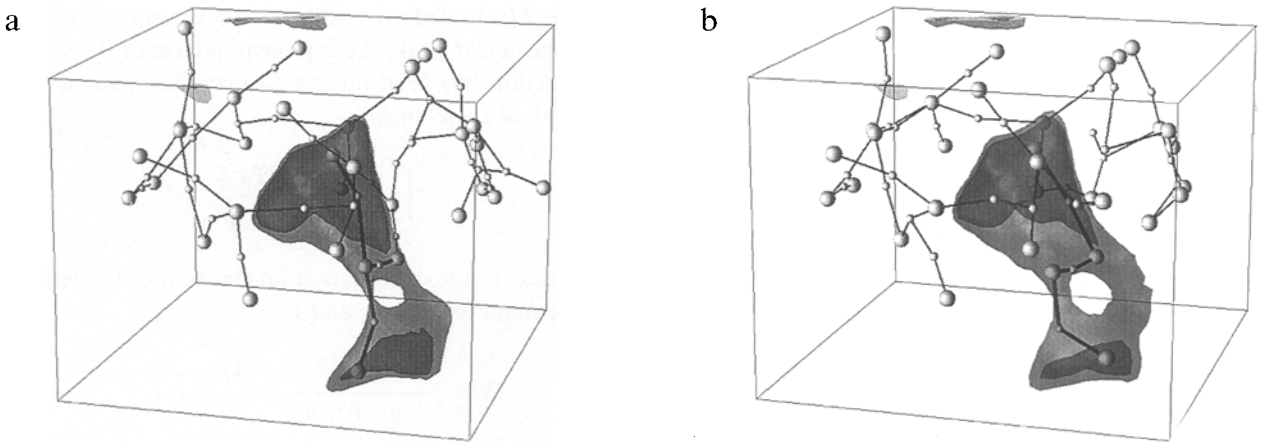

Fig. 4. Superimposition of the critical point representations (peaks: large balls; passes: small balls) of the cyclodextrin host and the (a) *R*- and (b) *S*-flurbiprofen guest obtained using ORCRIT on 3 Å reconstructed electron density maps. Isopotential energy contours, calculated using a probing sphere of radius 2.5 Å, are shown at −1.2 and −1.05 (in arbitrary units) for the *R*-complex and at −1.2 and −1.00 for the *S*-complex. The pictures were drawn using IBM Visualization DataExplorer, v. 1.2.

functions, these authors are able to derive gradients and Hessians of the nuclear coordinate derivatives, i.e., similar properties as critical point characteristics.

Shape complementarity analysis

The application of the program ORCRIT at 3 Å resolution showed that the cyclodextrin and the flurbiprofen were represented by 27 (system *S*) or 28 (system *R*), and 4 peaks, respectively [33]. The cyclodextrin–flurbiprofen complex systems are reduced to representations where peaks encompass specific sets of atoms: each glucose ring, CH_2OCH_3, and methoxy group. However, a few exceptions occur, such as for one of the cyclodextrin glucose rings involving two peaks. This peak is located close to the fluorine atom of the *R*- or *S*-flurbiprofen. Three methoxy peaks are missing, one for the *R*- and two for the *S*-complex. The flurbiprofen representation is similar in both cases: there is one peak for each phenyl ring, one for the fluorine atom, and the last peak represents the carboxylic group located outside the cyclodextrin cavity. The presence of the water molecule in the case of the *R*-flurbiprofen complex does not involve any additional peak.

Two different calculations of intermolecular steric interactions were carried out. First, the interaction energy of a sphere was probed at each node of a grid (grid interval = 0.35 Å) in order to characterize the sterically accessible volume within the cyclodextrin. Second, the steric interaction between all peaks of the flurbiprofen molecule and all peaks of the cyclodextrin hosts was calculated in order to evaluate the relative stabilization of one enantiomer with respect to the other. The interaction energy of a peak or a probing sphere at one particular location includes the contribution of all the host peaks and their periodic images corresponding to two unit cells on either side of the original unit cell, along each of the three spatial directions. The periodic images were obtained by considering the space-group symmetry operations and translations acting on the coordinates and the eigenvector orientations of each peak.

TABLE 3
MEAN DENSITY VALUES AND RADII ASSOCIATED WITH EACH PEAK OF THE CRITICAL POINT REPRESENTATIONS[a]

	R				*S*			
	ρ ($e^-/Å^3$)	r_x (Å)	r_y (Å)	r_z (Å)	ρ ($e^-/Å^3$)	r_x (Å)	r_y (Å)	r_z (Å)
Cyclodextrin								
Glucose rings	1.74±0.12	1.50±0.08	2.07±0.17	3.15±0.68	1.76±0.09	1.50±0.09	2.83±0.15	3.23±0.62
CH_2OCH_3	1.71±0.11	1.40±0.06	1.60±0.08	2.68±0.69	1.77±0.12	1.40±0.06	1.57±0.08	2.53±0.36
OCH_3	1.58±0.18	1.42±0.07	1.83±0.22	2.55±0.43	1.56±0.19	1.42±0.07	1.84±0.07	2.59±0.36
Flurbiprofen								
F	1.67	1.39	2.28	2.80	1.72	1.42	1.98	2.30
F-phenyl	1.64	1.39	2.76	3.92	1.63	1.37	2.25	2.75
Phenyl	1.57	1.45	1.91	3.82	1.43	1.62	1.80	3.68
COO	1.76	1.54	1.81	2.86	1.89	1.48	1.87	2.08

[a] The data were obtained using ORCRIT on reconstructed electron density maps of *R*- and *S*-flurbiprofen/cyclodextrin complexes at 3 Å resolution.

The intermolecular steric interaction potential was generated using a probing sphere, the radius of which was set to 2.5 Å, i.e., a value close to the largest guest peak radii defined along the x- and y-axes of the ellipsoids (Table 3). The r_z radii are defined along the critical point connections.

The obtained 3D steric isopotential contours are shown in Figs. 4a and b for the systems COYXAP10 and COYXET20, respectively. The steric interactions generated by the cyclodextrin cavity considering the crystal packing give rise to two main potential wells: one inside the cavity, and another outside the cavity. From the visual analysis of the 3D contours, it is observed that both flurbiprofen graphs are located within the minimal energy zones predicted by our method (Fig. 4). The accessible volume for *S*-flurbiprofen is apparently larger than the corresponding volume for *R*-flurbiprofen and allows the complete embedding of the flurbiprofen tree. There is also a higher potential energy area located at the level of the F-phenyl group.

In order to compare our results with a more conventional approach, we have determined the sterically accessible volume of each cyclodextrin host towards a probing sphere (r = 2.5 Å), considering the cyclodextrin atoms as vdW hard spheres. Similarly to our previous approach, a Lennard-Jones interaction energy was evaluated considering the volume of these vdW spheres. As H atoms were not reported in the CSD data files, CH, CH_2, and CH_3 groups were represented using vdW spheres with a radius of 2 Å. In order to remain consistent with the host representation level, the radius of the probing sphere was also defined in the atomic resolution range. A value of 1.4 Å, a typical value for oxygen atoms, was selected. The accessible volume thus obtained is displayed in Fig. 5 for each complex by means of isopotential energy levels plotted at −0.8 and −0.6 (in arbitrary units). The figures show that the contour shapes have some patterns similar to those determined at a resolution of 3 Å, but are characterized by slightly higher energy values. Again, a higher potential energy zone is located in the area of the F-phenyl group of the flurbiprofen molecule. Both flurbiprofen configurations can be accommodated by the cyclodextrin host. Representations of the accessible volume are therefore similar both at high- and medium-resolution levels. The electron density function is thus a very adequate tool to study shape complementarity, not only at atomic resolution, but also at lower levels. The local information retained at critical point locations (density, gradient, curvature) is sufficient to define accessible volume. It is very interesting to note that both high- and medium-resolution potential energy functions present similar topological features. A topological analysis of the potential energy function itself, using for example Mezey's approach [5], should tell which topological features are retained when moving from one resolution to another.

Finally, the total steric interaction potential energy values between the *R*- and *S*-flurbiprofen molecules and the cyclodextrin were computed using Eqs. 11–14. The values obtained were −3.95 and −2.87 (arbitrary units), respectively. Steric interactions thus slightly favour the stabilization of the *R*-flurbiprofen. In order to validate these interaction energy values, comparisons were made with values available in the literature. A paper by Lipkowitz et al. [34] reported energy minimizations using the AMBER potential function for a system composed of three cyclodextrin structures and one flurbiprofen molecule. The atoms of the cyclodextrins were held fixed at their crystallographic locations, while the guest molecule was allowed to relax. The authors found that complexation is energetically favored for the *S*-flurbiprofen with respect to the *R*-form. This result is opposite to our own observations, and additional investigations were thus carried out in order to explain the trend followed by our electron density topological approach [33]. These calculations were carried out using a conventional molecular mechanics energy minimization procedure, which was

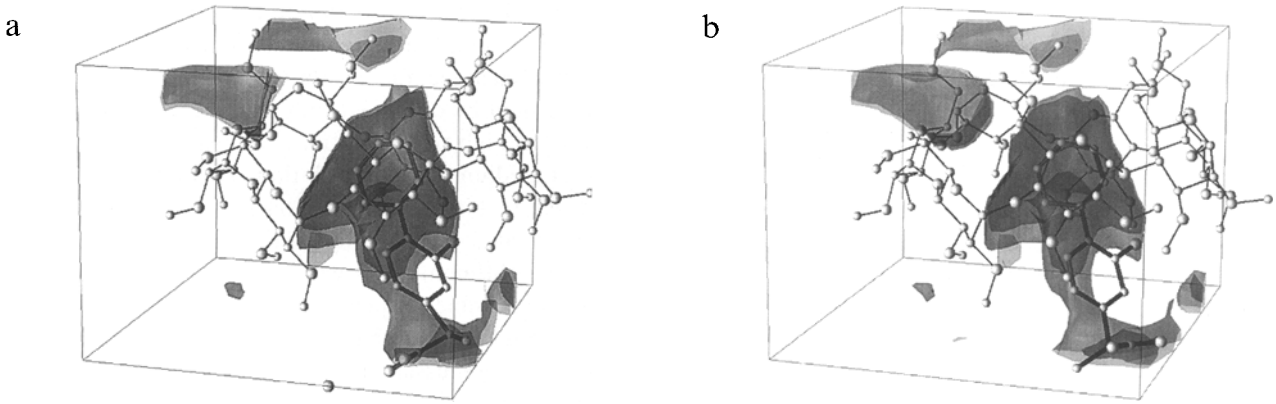

Fig. 5. Superimposition of the molecular structures of the cyclodextrin host and the (a) *R*- and (b) *S*-flurbiprofen guest obtained from the CSD. Isopotential energy contours were calculated using a sphere of radius 2.5 Å (contours are shown at −0.8 and −0.6 in arbitrary units), probing a vdW representation for the cyclodextrin host through a Lennard-Jones potential energy evaluation. The pictures were drawn using IBM Visualization DataExplorer, v. 1.2.

applied to each flurbiprofen/β-cyclodextrin complex under various conditions using Discover, v. 94.0 [35]. The hydrogen atoms, which were not present in the CSD entry, were added using InsightII [36]. In this operation, the unfilled valences were completed according to the hybridization state and the geometry of the unsaturated atom and its surrounding environment.

The minimization scheme involved three successive steps: first, a steepest descent algorithm was applied, which was carried out with a convergence threshold of 100 kcal mol^{-1} $Å^{-1}$. Then, at convergence of the first step, the process switched to a truncated Newton and Polak–Ribiere conjugate gradient optimization method, which was stopped when the gradient rms reached the threshold of 10 kcal mol^{-1} $Å^{-1}$. The last stage consisted in a Broyden–Fletcher–Goldfarb–Shanno algorithm applied with a convergence limit of 0.001 kcal mol^{-1} $Å^{-1}$. Three sets of minimization conditions were selected. In a first approach, conditions similar to those reported in Lipkowitz's paper [34] were considered, i.e., the energy of the guest molecule is minimized in a rigid environment made of three cyclodextrin structures. As specified by Lipkowitz et al., the dielectric of the medium was set to 1.5 and no cutoffs were used. During the minimization procedure, all hydrogen atoms were relaxed. In a second approach, conditions identical to those described above were selected and one water molecule was added at its crystallographic coordinates to the *R*-complex structure. In a last approach, the crystal structure of the host–guest systems was fully considered. An Ewald summation method was used to take into account the long-range electrostatic interactions. All atoms of the host–guest system, except the hydrogens, were kept fixed at their crystallographic locations previously used to generate the EDMs. The water molecule was included in the *R*-system structure. Discover [35] was used with the so-called cff91 potential [37]. The binding enthalpy ΔH_{bind}, calculated for each flurbiprofen configuration, was determined by subtracting from the energy of the host–guest complex the energy of the guest and the host infinitely separated from each other by holding the internal degrees of freedom fixed for both molecules at their predetermined values:

$$\Delta H_{bind} = E_{complex} - (E_{guest} + E_{host}) \qquad (15)$$

The tendency reported by Lipkowitz was observed, i.e., the binding enthalpy ΔH_{bind} is lower for the *S*-complex when the first set of minimization conditions is used: $\Delta H_{bind}(R) = -44.16$ and $\Delta H_{bind}(S) = -45.64$ kcal mol^{-1}. However, this trend is reversed when a water molecule and/or periodicity are considered, as already predicted by the method based on the critical point size and location. The trend is indeed slightly reversed when the water molecule is present: $\Delta H_{bind}(R) = -48.36$ and $\Delta H_{bind}(S) = -45.64$ kcal mol^{-1}; while there is a large reversal when the crystal periodicity is considered: $\Delta H_{bind}(R) = -218.83$ and $\Delta H_{bind}(S) = -193.95$ kcal mol^{-1}. Under such conditions we obtained, as observed with the ORCRIT method, a stronger stabilization of the *R*-complex.

Conclusions

We have described a topological analysis approach that converts an electron density map into its minimal spanning tree of critical points. One of the aims of a topological analyis is to provide a brief characterization of the molecular shape. Two methods were derived, which allow (i) the comparison through similarity evaluation of a critical point tree of proteins and secondary structure models, and (ii) the evaluation of accessibility of a guest molecule into a receptor through the calculation of a potential function, which is seen as a signature of the fit between the two species.

First, the segmentation and successful interpretation of the segmented parts of protein electron density maps validate the ability of the critical point approach to represent structural information at medium resolution. The local information characterizing each critical point is also sufficient to determine an adequate connectivity pattern. It is important to keep in mind that these conclusions are valid for maps reconstructed without phase error. Under these conditions, a 3 Å resolution will segment a protein structure into its residues and their side chains. The secondary structure is determined from geometric deviations of the critical point segments with respect to templates modelled from a statistical analysis of PDB structures.

Second, our very simple intermolecular energy calculations, which are based on the location and volume of local maxima of an electron density function, showed that the *R*-flurbiprofen structure is sterically more stabilized in the cyclodextrin host than the *S*-flurbiprofen structure. This is also observed when calculating the binding enthalpy using a conventional molecular mechanics method. The application of such an approach as a representation of volumic information is thus validated. However, from a geometric point of view chiral molecules cannot be differentiated at medium resolution, since the chiral centre is represented by one point only. Indeed, objects that are different at high resolution may look similar when compared at medium or low resolution. This was emphasized by Mezey in his paper on similarity of 3D bodies [38]. Therefore, at 3 Å resolution, chiral compounds cannot always be differentiated, while at atomic resolution it should be possible to quantify their dissimilarity [39].

The consideration that molecular structures can be described by Gaussian functions both at medium and high resolution leads us to think that hierarchical representations are possible through the use of ORCRIT at various resolution levels, or through the use of a wavelet decomposition procedure as described by Butzlaff et al.

[25] and Marr, as reported in Ref. 40. Decomposition of topological representations into wavelet expansion coefficients is now under investigation.

Acknowledgements

The authors wish to thank Prof. J. Glasgow and Prof. S. Fortier of the Queen's University (Kingston, Canada) and Dr. F.H. Allen (Cambridge Crystallographic Structural Data Centre, England) for fruitful collaboration. They also thank C.K. Johnson for providing the ORCRIT program, and the FUNDP for the use of the Namur Scientific Computing Facility (SCF) Center. They acknowledge the financial support of the FNRS-FRFC, the 'Loterie Nationale' for convention no. 9.4593.92, the FNRS within the framework of the 'Action d'impulsion à la recherche fondamentale' of the Belgian Ministry of Science under convention no. D.4511.93, IBM Belgium for the Academic Joint Study on 'Cooperative Processing for Theoretical Physics and Chemistry', and NSERC for a Collaborative Grant. T.L. thanks the 'Services de la Programmation Scientifique' (SPPS) for his grant in the framework of the ELSAM (Electronic Large Scale Computational System for Advance Materials) project, part of the Belgian National Program of 'Impulsion in Information Technology'. L.L. thanks the FNRS for scientific research support and for her 'Chargé de Recherches' position.

References

1 a. Johnson, C.K., In Proceedings of the American Crystallographic Association Meeting, Evanston, IL, 1976, abstract B1.
b. Johnson, C.K., In Proceedings of the American Crystallographic Association Meeting, Asilomar, CA, 1977, abstract JQ6.
c. Johnson, C.K., ORCRIT: The Oak Ridge Critical Point Network Program, Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, TN, 1977.
2 Terry, A., Ph.D. Thesis, Stanford University, Stanford, CA, 1983.
3 Leherte, L., Fortier, S., Glasgow, J.I. and Allen, F.H., Acta Crystallogr., D50 (1994) 155.
4 Leherte, L. and Allen, F.H., J. Comput.-Aided Mol. Design, 8 (1994) 257.
5 a. Mezey, P.G., J. Comput. Chem., 8 (1987) 462.
b. Mezey, P.G., In Johnson, M.A. and Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, Wiley, New York, NY, 1990, pp. 321–368.
6 a. Hodgkin, E.E. and Richards, W.G., Int. J. Quantum Chem., Quantum Biol. Symp., 14 (1987) 105.
b. Good, A.C., J. Mol. Graphics, 10 (1992) 144.
c. Cioslowski, J. and Fleishmann, E.D., J. Am. Chem. Soc., 113 (1994) 64.
d. Mestres, J., Sola, M., Duran, M. and Carbo, R., J. Comput. Chem., 15 (1994) 1113.
7 a. Greer, J., J. Mol. Biol., 82 (1974) 279.
b. Greer, J., J. Mol. Biol., 100 (1976) 427.
c. Jones, T.A., Zou, J.Y., Cowan, S.W. and Kjeldgaard, M., Acta Crystallogr., A47 (1991) 110.
8 Marr, D., Vision, W.H. Freeman, San Francisco, CA, 1982.
9 a. Richard, F.M., Annu. Rev. Biophys. Bioeng., 6 (1977) 151.
b. Connolly, M.L., J. Appl. Crystallogr., 16 (1983) 548.
10 Del Carpio, C.A., Takahashi, Y. and Sasaki, S.-I., J. Mol. Graphics, 11 (1993) 23.
11 Santavy, M. and Kypr, J., J. Mol. Graphics, 2 (1984) 47.
12 Arteca, G.A. and Mezey, P.G., J. Comput. Chem., 9 (1988) 554.
13 Lin, S.L., Nussinov, R., Fischer, D. and Wolfson, H., Protein Struct. Funct. Genet., 18 (1994) 94.
14 a. Leicester, S., Finney, J. and Bywater, R., J. Math. Chem., 16 (1994) 315.
b. Leicester, S., Finney, J. and Bywater, R., J. Math. Chem., 16 (1994) 343.
15 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
16 Grootenhuis, P.D.J., Rae, D.C., Kollman, P.A. and Kuntz, I.D., J. Comput.-Aided Mol. Design, 8 (1994) 731.
17 a. Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.
b. Allen, F.H., Kennard, O. and Taylor, R., Acc. Chem. Res., 16 (1983) 146.
18 Smith, V.H., Price, P.F. and Absar, I., Isr. J. Chem., 16 (1977) 187.
19 Howard, S.T., Hursthouse, M.B., Lehman, C.W., Mallinson, P.R. and Frampton, C.S., J. Chem. Phys., 97 (1992) 5616.
20 Bader, R.W., Atoms in Molecules – A Quantum Theory, Clarendon, Oxford, 1990.
21 Johnson, C.K., In Proceedings of the American Crystallographic Association Meeting, Pittsburgh, PA, 1992, abstract PA99.
22 Popelier, P.L.A., Theor. Chim. Acta, 87 (1994) 465.
23 a. Shirsat, R.N., Bapat, S.V. and Gadre, S.R., Chem. Phys. Lett., 200 (1992) 373.
b. Gadre, S.R. and Shrivactava, I.H., Chem. Phys. Lett., 204 (1993) 350.
24 a. Glasgow, J.I., Fortier, S. and Allen, F.H., In Hunter, L. (Ed.) Artificial Intelligence and Molecular Biology, AAAI Press, Menlo Park, CA, 1993, pp. 433–458.
b. Fortier, S., Castleden, I., Glasgow, J.I., Conklin, D., Walmsley, C., Leherte, L. and Allen, F.H., Acta Crystallogr., D49 (1993) 168.
25 Butzlaff, M., Dahmen, W., Diekmann, S., Dress, A., Schmitt, E. and von Kitzing, E., J. Math. Chem., 15 (1994) 77.
26 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
27 Hall, S.R. and Stewart, J.M. (Eds.) Xtal 3.0 Reference Manual, Universities of Western Australia, Nedlands, and Maryland, College Park, MD, 1990.
28 Leherte, L., Baxter, K., Glasgow, J.I. and Fortier, S., In Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D., Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, MIT/AAAI Press, Menlo Park, CA, 1994, pp. 261–268.
29 Rich, E. and Knight, K., Artificial Intelligence, McGraw-Hill, New York, NY, 1991.
30 Patterson, D.W., Introduction to Artificial Intelligence and Expert Systems, Prentice Hall, Englewood Cliffs, NJ, 1990.
31 Hutchinson, S.A., Cromwell, R.L. and Kak, A.C., In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, 1989, pp. 541–548.
32 Grant, J.A. and Pickup, B.T., J. Phys. Chem., 99 (1995) 3503.
33 Leherte, L., Latour, Th. and Vercauteren, D.P., Supramol. Sci., in press.

34 Lipkowitz, K.B., Green, K.M., Yang, J.-A., Pearl, G. and Peterson, M.A., Chirality, 5 (1993) 51.
35 Discover, User Guide Parts I, II and III, v. 94.0, Biosym Technologies, San Diego, CA, 1994.
36 InsightII, User Guide Parts I and II, v. 2.3.0, Biosym Technologies, San Diego, CA, 1993.
37 a. Maple, J.R., Dinur, U. and Hagler, A.T., Proc. Natl. Acad. Sci. USA, 85 (1988) 5350.
b. Waldman, M. and Hagler, A.T., J. Comput. Chem., 14 (1993) 1077.
c. Maple, J.R., Hwang, M.J., Stockfish, T.P., Dinur, U., Waldman, M., Ewig, C.S. and Hagler, A.T., J. Comput. Chem., 15 (1994) 162.
38 Mezey, P.G., J. Math. Chem., 7 (1991) 39.
39 a. Meyer, A.Y. and Richards, W.G., J. Comput.-Aided Mol. Design, 5 (1991) 427.
b. Zabrodsky, H. and Avnir, D., J. Am. Chem. Soc., 117 (1995) 462.
40 Meyer, Y., Wavelets – Algorithms & Applications, SIAM, Philadelphia, PA, 1993.