Journal of Computer-Aided Molecular Design (2005) 19: 525–549
DOI 10.1007/s10822-005-9005-3

Influence of conformation on the representation of small flexible molecules at low resolution: alignment of endothiapepsin ligands

Laurence Leherte*, Nathalie Meurice† & Daniel P. Vercauteren
Laboratoire de Physico-Chimie Informatique (PCI), Facultés Universitaires Notre-Dame de la Paix (FUNDP), Rue de Bruxelles 61, B-5000, Namur, Belgium

Received 9 December 2004; accepted 23 July 2005

Key words: Electron density distribution, critical points, endothiapepsin ligands, flexible molecules, molecular fragments, molecular alignment

Summary

In this contribution, we discuss a molecular representation mode for the generation of reduced descriptions of flexible molecules in various conformations. The representations of the endothiapepsin ligands are constituted by graphs of peaks obtained through a hierarchical merging algorithm which combines the location of promolecular electron density (ED) maxima with the decomposition of the molecular structures into fragments. The representations are then aligned through the use of a Monte Carlo/Simulated Annealing procedure. The evaluation function of the alignment solutions is based on the local density values of the peaks and their inter-distances. The applications show that the alignment of a given molecule onto itself, in a different conformation, is successful when a pseudo-topological path length is considered in the evaluation function of the solution, while Cartesian distances are more adapted to the scoring for alignments of two different molecules in their co-crystallized conformation. Results are compared with the available literature data.

Introduction

In drug design applications, it is common to build new potent molecules from the available structural information of their specified biological receptor. However, the three-dimensional (3D) structure of a receptor is not always precisely known. In such a case, alternative approaches, such as molecular alignments of two or more ligands for a given biological receptor, need to be considered. The most basic comparison approaches usually require the knowledge of the atom types and their connectivity. The use of 2D representations overcomes the more time consuming adjustment procedures of relative 3D molecular orientations [1], especially when matching 3D grids of points, e.g., electron density (ED) distribution functions or molecular electrostatic potential (MEP) maps. Several molecular similarity evaluation approaches are thus based on refined molecular field matching techniques. For example, similarity quantifiers applicable to steric volume and electrostatic fields have been implemented in the program MIMIC [2]. This program calculates the so-called Carbó index [3], one of the most widely used molecular similarity measure between molecules. Orientation-independent descriptors possess the obvious advantage of avoiding molecular translation and reorientation operations when matching two or more molecular structures. In this context, reference studies were achieved by Mezey [4] through his Shape Group Method which

*To whom correspondence should be addressed. Phone: +32-81-72-45-60; Fax: +32-81-72-54-66; E-mail: laurence.leherte@fundp.ac.be
†F.N.R.S. Scientific Research Worker

allows the characterization of molecular surfaces (van der Waals envelopes, iso-density contours, MEP iso-contours, ...) and their partitioning into domains using curvature, a topology-related information. Bader [5] established a topological analysis method of 3D ED distributions in terms of the number and kind of their critical points (CP), i.e., the points where the gradient of the density is equal to zero. Popelier [6] has later extended this approach to include the concept of molecular similarity. The author proposed to consider the properties of bond critical points (BCP) extracted from quantum mechanical *ab initio* wavefunctions. MEP functions have also been the subject of topological/topographical studies [7, 8]; these were for example used by Willett and coworkers [9, 10] in order to generate the so-called field-graphs. Such graph representations facilitate the alignment of MEP fields through genetic algorithms (GAs). This is particularly important for similarity search in large databases where speed and efficiency are always simultaneously required.

In previous works of our group [11–15], both GA and Monte Carlo/Simulated Annealing (MC/SA) algorithms were applied to the alignment of small biomolecules in a given conformation using topographical information extracted from their ED distribution. The particularity of our works resides in the use of low resolution ED distributions to generate reduced graph representations, i.e., representations above the atomic level.

In the present work, we investigate the influence of conformational changes on graph representations and alignment results obtained using promolecular ED data at various degrees of smoothing. Glick et al. have already explored the representation of flexible molecules using descriptions at several scales [16, 17]. In their work, a hierarchy of models generated using a *k*-mean clustering algorithm for the ligand under consideration was established starting from the lowest resolution representation of the ligand, i.e., one single point located at the mean position of the ligand atoms.

In the literature, one finds that the conformational flexibility of a molecule can be modeled by a discrete set of preferred geometrical parameters (e.g., torsion angles, ring conformations, ...) issued from databases of structural information, such as CSD [18], or from low-energy conformations generated, for example, through molecular dynamics procedures [19, 20], or knowledge-based systems [21–23]. However, the use of a finite number of conformations does not guarantee that the active conformation of a molecule is actually present in the given discrete list. Thorner et al. [10] indeed showed that MEP-based searching is possible only if molecules are represented by a large number of different conformational graphs. But it is sometimes necessary to limit the number of conformations, for example by keeping only energetically acceptable conformations [24], or by generating template structures [24–27]. Another treatment of molecular flexibility consists in the consideration of a continuous set of conformations, for example through the definition of distance ranges [28–31]. Also, Raymond and Willett [32] implemented the adjustment of distance ranges in order to verify triangle and tetrangle inequalities, and Mills et al. [33] proposed to use simulated annealing to minimise the difference between the distance matrices calculated from the so-called "critical points" of two ligands, i.e., hydrogen-bond and aromatic ring "positions".

Molecular flexibility can also be treated by considering, as above, only one initial conformation, but allowing its geometrical parameters to be modified during the alignment or docking process [34–39]. In those cases, molecular geometry is often directly implied in the calculation of a scoring index through molecular mechanics type formulae [22, 37, 38], or feature-based formulae [36] such as volume, aromaticity, H-bond donor/acceptor, molecular refractivity, surface exposure, ... Klebe et al., for example, included a forcefield type expression into the alignment function of the program SEAL [22]. This extension allowed the comparison and alignment of strongly different molecular skeletons without the need of preliminary assumptions about their respective orientation. Molecular flexibility can also be addressed through the approach of molecular fragmentation [40–43], which implies the decomposition of a molecule into fragments and its further reassembling. An incremental construction procedure wherein a molecule is iteratively reconstructed to best fit a reference structure can be found in the program FlexS, a derivative of FlexX, as well as in the program DOCK4.0 [44–46].

In the next section, we describe the theoretical concepts that are used in this work, i.e., the definition of a promolecular ED representation,

the smoothing technique through the diffusion problem, and the decomposition algorithm which allows assignment of the resulting molecular fragments to the local maxima of an ED distribution. Then, decomposition results are presented and discussed for a family of five highly flexible endothiapepsin ligands considered in various conformations. Alignments obtained using a MC/SA procedure are finally presented at different degrees of smoothing and results are compared with literature data.

Theoretical background

Promolecular atomic shell approximation

Promolecular models have often been shown as reasonable or even very good approximation levels to model electron density (ED) distributions, for example in chemical bond analysis or molecular similarity applications [47–56] even if the ED topology may differ between experimental and theoretical models especially at covalent bonds [56]. Analytical descriptions of promolecular ED distributions are either based on atomic or ionic wavefunctions [47–49, 52, 56], exponential functions [51], or fitted Gaussian functions [50, 52, 55]. In the Promolecular Atomic Shell Approximation (PASA) approach, a promolecular ED distribution ρ_{M} is analytically represented as a weighted summation over atomic ED distributions ρ_{a}, which are described in terms of series of squared 1s Gaussian functions fitted from atomic basis set representations [57]:

$$\rho_{\mathrm{a}}(\mathbf{r}-\mathbf{R}_{\mathrm{a}}) = \sum_{i=1}^{5} w_{\mathrm{a},i}\left[\left(\frac{2\varsigma_{\mathrm{a},i}}{\pi}\right)^{3/4} e^{-\varsigma_{\mathrm{a},i}|\mathbf{r}-\mathbf{R}_{\mathrm{a}}|^2}\right]^2 \tag{1}$$

where $\mathbf{R}_{\mathrm{a}}$ is the position vector of atom a, and $w_{\mathrm{a},i}$ and $\zeta_{\mathrm{a},i}$ are the fitted parameters, respectively, as reported in reference [58]. ρ_{M} is then calculated as:

$$\rho_M = \sum_{\mathrm{a}} Z_{\mathrm{a}}\rho_{\mathrm{a}} \tag{2}$$

where Z_{a} is the atomic number of atom a and $\sum_{i=1}^{5} w_{\mathrm{a},i} = 1$.

In the present approach to generate smoothed three-dimensional (3D) ED functions, ρ_{M} is directly expressed as the solution of the diffusion equation according to the formalism presented by Kostrowicki et al. [59]:

$$\rho_{\mathrm{a},t}(\mathbf{r}-\mathbf{R}_{\mathrm{a}}) = Z_{\mathrm{a}}\sum_{i=1}^{5}\alpha_{\mathrm{a},i}(1+4\beta_{\mathrm{a},i}t)^{-3/2}e^{\frac{-\beta_{\mathrm{a},i}|\mathbf{r}-\mathbf{R}_{\mathrm{a}}|^2}{1+4\beta_{\mathrm{a},i}t}} \tag{3}$$

where

$$\beta_{\mathrm{a},i} = 2\varsigma_{\mathrm{a},i} \text{ and } \alpha_{\mathrm{a},i} = w_{\mathrm{a},i}\left(\frac{\beta_{\mathrm{a},i}}{\pi}\right)^{3/2} \tag{4}$$

In previous works, we applied the program ORCRIT [60] to generate critical point (CP) graph representations of molecules from their smoothed ED distributions, and we used the resulting graph representations for molecular alignment purposes [11–15]. These graphs consisted in a set of CPs located in molecular ED that were calculated at medium crystallographic resolution, or that were smoothed using various mathematical tools such as wavelet analysis or by application of the diffusion equation. Even if all these approaches are conceptually different, they all lead to alike reduced representations. In a given ED map, all CPs were considered connected to each other, i.e., they formed a fully connected graph. The considered molecular descriptors were the density value of the CPs and the distances between the CPs. In combination with GA and MC/SA superposition algorithms, we showed how pharmacological molecules with a similar property could be superimposed, and hence determine the corresponding chemical groups between the molecules. In these previous works, the chemical groups were defined on a visual basis only. However, as the approach appeared to work well in the determination of pharmacophore elements, a further step was to elaborate a method to associate to each CP, and more precisely to each local maxima of an ED distribution, a precise chemical content. This led us to the implementation of a method to provide a chemical meaning to the low resolution peaks [61, 62]. In the present paper, we apply this method to the generation of representations of molecules that can adopt very different conformational states, which was not the case in the previous papers. We later use our superposition technique to evaluate a simple objective function (a) in the recognition of molecular fragments with similar atom content, and (b) in the alignment of molecules with variable

chemical content but with a similar conformation. In the last case, several resolution levels are considered to determine the ED smoothing levels that carry useful information to align molecular structures.

Hierarchical merging/clustering algorithm

In order to follow the pattern of local maxima in a molecular ED distribution, as a function of the degree of smoothing, we implemented the algorithm described by Leung et al. [63]. The various steps of the resulting merging/clustering algorithm are:

1. At scale $t = 0$, each atom of a molecular structure is considered as a local maximum (peak) of the promolecular ED distribution function. All atoms are thus considered as the starting points of the merging procedure described below.
2. As t increases from 0.0 to a given maximal value t_{max}, the ED distribution topology is modified. Each peak thus moves continuously along the gradient path to reach a location in the 3D space where $\nabla\rho = 0$. On a practical point of view, this consists in following the trajectory of the peaks obtained at a resolution t on the ED distribution surface calculated at resolution level $t + \Delta t$:

$$\mathbf{r}_{peak}(t + \Delta t) = \mathbf{r}_{peak}(t) + \frac{\Delta}{\rho_{peak}(t)} \nabla\rho_{peak}(t + \Delta t) \tag{5}$$

The trajectory search is stopped when $\nabla\rho_{peak}(t)$ is lower or equal to a limit value $grad_{lim}$. Some trajectories may have converged to a precise location. In such a case, the convergence is considered effective when the associated peaks are separated by a distance that is lower than the initial value of $\Delta^{1/2}$. One thus considers that there is only one peak left. The procedure is repeated for each selected value of t. If the initial Δ value is too small to allow convergence towards a local maximum of the ED within the given number of iterations, its value is doubled (a scaling factor that is arbitrarily selected) and the procedure is repeated until final convergence.

A description about how Equation (5) was adapted to the study of molecular ED distributions was previously reported [61]. It was successfully applied to small molecules containing the following elements: H, C, N, O, S, Cl [61], and to proteins [62]. In the last paper, it is additionally showed that t is proportional to the well-known crystallographic isotropic temperature factor. In all cases treated so far, Δ and the number of iterations were always set equal to 0.0001 bohr^2 and 2000, respectively. The value of $grad_{lim}$ may however be adapted depending upon the size of the molecules and the need to get precise values for the second derivatives of the ED at the local maxima. In the present work on small molecules, $grad_{lim}$ was set equal to a value of 0.00001 e$^-$/bohr^4.

When combined with Equation (3), the above proposed decomposition procedure does not require any calculation of the ED maps of the molecules. The results obtained using the present algorithm are the location of the local maxima (peaks), their density, eigenvectors, and Laplacian values, as well as the atomic content of all fragments, at each value of t between 0 and t_{max}. The t-dependent clustering can be represented using dendrograms. Visual results are generated using the Web version of the program Phylodendron [64]. Input data were written in the adequate format using DENDRO [65], an in-house code implemented using Delphi, an object-oriented programming language that allows representation and processing of data in terms of classes of objects. That approach is thus inverse of the one presented by Glick et al. [16, 17] already cited above.

Superposition algorithm

As already mentioned, we used the procedure that was already applied previously [13–15], i.e., a MC/SA method for the pairwise matching of the molecular CP graphs obtained from the molecular structures at various resolutions. Molecular conformers are represented using local ED properties, i.e., the ED values associated with the peaks, and their inter-distances. As there is no knowledge about the intramolecular forces that govern a graph conformation, molecular flexibility could only be approached by considering a limited set of different rigid conformers as described in the Introduction. The MC algorithm is intended to generate a sequence of solutions satisfying a

predetermined probability distribution function. Those solutions are selected among all possibilities using random numbers. During a Metropolis MC process, only sufficiently probable solutions are kept in order to favor convergence toward the selected probability distribution [66]. The basic form of the Metropolis MC algorithm can be briefly described as a sequence of the following steps: generate an initial solution consisting of a random matching between the CPs of two molecules, score the initial solution, generate a new solution and evaluate it, use a Boltzmann-based probability function p to determine whether the new solution is retained or not, repeat the previous steps until the desired number of matchings is obtained. The SA simulation part [67] consists, in the present work, in considering a sequence of several MC sampling procedures that are carried out at progressively decreasing rates of acceptance:

$$p = e^{-\gamma \mathrm{RMS}} \tag{6}$$

where γ is the parameter that controls the rate of acceptance and modulates the evaluation function $\mathrm{RMS_{AB}}$ between molecules A and B:

$$\mathrm{RMS_{AB}} = \sqrt{\frac{w_\rho}{n}\sum_{i=1}^{n}\left(\rho_i - \rho_i^{ref}\right)^2 + \frac{w_d}{\mathrm{nbn}}\sum_{i=1}^{\mathrm{nbn}}\left(d_i - d_i^{ref}\right)^2} \tag{7}$$

where, in a completely connected CP graph composed of n points, the number of connections between the points is given by $\mathrm{nbn} = n(n-1)/2$ and w_ρ and w_d are the weights given to the density and distance contributions, respectively, and are set equal to 1. In Equation (7), the reference graph is the smallest one, i.e., either graph of molecule A or B. The SA approach allows, first, to accept numerous matches, even those with low probabilities, and then, as γ decreases, to progressively limit the acceptance of solutions to the most probable ones. This technique is often used in optimization problems characterized by many local solutions, e.g., in conformational analysis.

The above proposed $\mathrm{RMS_{AB}}$ expression however does not provide well-defined limits to the similarity degree between two molecules. Indeed, if two molecules are identical, $\mathrm{RMS_{Abmin}} = 0$, while its maximum stays undetermined if they are not identical. If one would discretize the well-known Carbó formula, for instance, a similarity index could thus be defined as:

$$S_{\mathrm{AB}} = S_\rho + S_d = \frac{\sum_{k=1}^{n}\rho_i\rho_j}{\sqrt{\sum_{k=1}^{n}\rho_i^2}\sqrt{\sum_{k=1}^{n}\rho_j^2}} + \frac{\sum_{l=1}^{\mathrm{nd}} d_i d_j}{\sqrt{\sum_{l=1}^{\mathrm{nd}} d_i^2}\sqrt{\sum_{l=1}^{\mathrm{nd}} d_j^2}} \qquad i \in A, j \in B \tag{8}$$

where n is the number of pairs of the matched CPs, and nd is the number of distances involving i (or j) with any of the other $n-1$ peaks within the single graph A (or B). Thus, the number of distances to be computed between a CP and its neighbors in the same graph is $\mathrm{nd} = n - 1$. S_{AB} has a maximal value of 2 and a theoretical lower limit of 0. In the Application part of the present paper, molecular alignments were obtained using both $\mathrm{RMS_{AB}}$ and S_{AB} as evaluation functions, but we observed that the final results were generated with a slower convergence rate and a lower degree of success when using S_{AB}. We explain this by the fact that, even if S_{AB} is not strictly correlated to $\mathrm{RMS_{AB}}$, there is a global decrease of $\mathrm{RMS_{AB}}$ which, close to convergence, gets stronger than the corresponding increase of similarity S_{AB} (Figure 1).

In the applications described herein, the starting solutions do not need to be close to the optimal solution. Also, they do not require any predetermined anchor elements between the reference and the fitted graphs. The reference molecule is always the smallest ED graph, as opposed to the program FlexS [44, 45]. Each SA run consisted of 20 MC procedures (each of 200000 iterations) that were carried out with γ ranging between 0.005 and 0.1 (for standardized data) or 0.05 and 1.0 (for unstandardized data). With S_{AB}, the following γ limits were always considered: 0.0001 and 0.05. Various superposition criteria were used, as detailed in the Application section. One of them consisted to standardize separately the density and distance values involved in the $\mathrm{RMS_{AB}}$ expression to avoid two contributions with different units and magnitudes. In the standardization procedure, each data a of a data set was transformed according to:

$$a_{\mathrm{st}} = \frac{a - \bar{a}}{\sigma_\mathrm{a}} \tag{9}$$

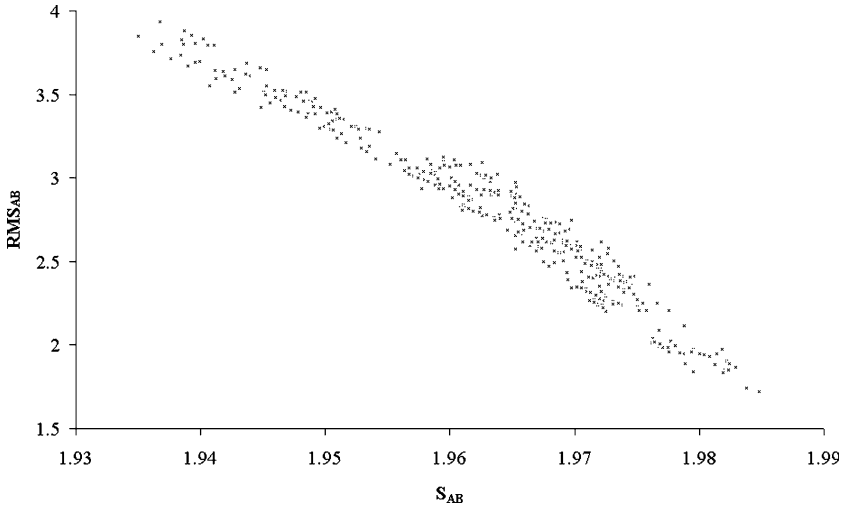

Figure 1. Dependence of the RMS_{AB} objective function with the corresponding S_{AB} similarity index. Values were calculated during the alignment procedure of structures *1_pdb* and 1_opt at $t = 1.50$ bohr^2.

where $\bar{a}$ and σ_a are the mean and the standard deviation of the data set, respectively.

Best peak matching results were verified by an additional MC run of 2×10^6 iterations with a γ value selected in order to obtain a number of accepted moves that is close to the number of rejected moves. Matching results were, when appropriate, translated into visual molecular superimpositions using the program QUATFIT [68] as follows. In a first stage, starting from the pairs of matched CPs as generated by our superposition procedure, a graph is superimposed onto its reference CP graph and the resulting translation vector and rotation matrix are stored. These two quantities are then applied to the 3D atom coordinates of the corresponding molecule.

Applications

Materials

In previous applications of low resolution molecular representations to the comparison of small molecules, we considered rigid benzodiazepine-related structures and thrombin inhibitors [11–15]. Here, a new set of five highly flexible endothiapepsin ligands was selected after the very interesting work of Lemmen et al. [45]. Endothiapepsin is a single chain proteinase of 330 amino acids and has a molecular weight of 33.8 kDa. The structure is largely of β-sheet type and consists of two related lobes of approximately 170 amino acids each. The active site resides in a pronounced cleft between the lobes.

Inhibitors based on synthetic and naturally occurring analogues have been shown, by X-ray crystallography, to bind in the active site cleft in extended conformations. A detailed comparison of the X-ray structures of 21 inhibitor complexes is given by Bailey and Cooper [69]. The hydrogen bonds which position the inhibitor main chain in the active site cleft are largely conserved from one inhibitor to another implying that the largest determinants of specificity are the van der Waals contacts between the enzyme and the ligand side chains. There are strongly conserved binding interactions at the inhibitor sites named P_4, P_3, P_1, $P_{1'}$, and $P_{2'}$, compared to the weaker binding and unfavorable geometry of the P_2 residue [69]. Side chains of the inhibitors can adopt different conformations to compensate for greater or lesser occupation of the neighboring residues.

The three-dimensional (3D) atomic coordinates of the five considered ligands are derived from the crystallographic data stored in the PDB files named *2er7*, *4er1*, *4er2*, *5er1*, and *5er2* [70]. The primary sequence and structure of these five peptidic ligands, respectively named H-261, PD-125967, Pepstatin A, BW624, and CP-69799, are shown in Figure 2. For convenience, the five molecules will be numbered *1* to *5* further in the

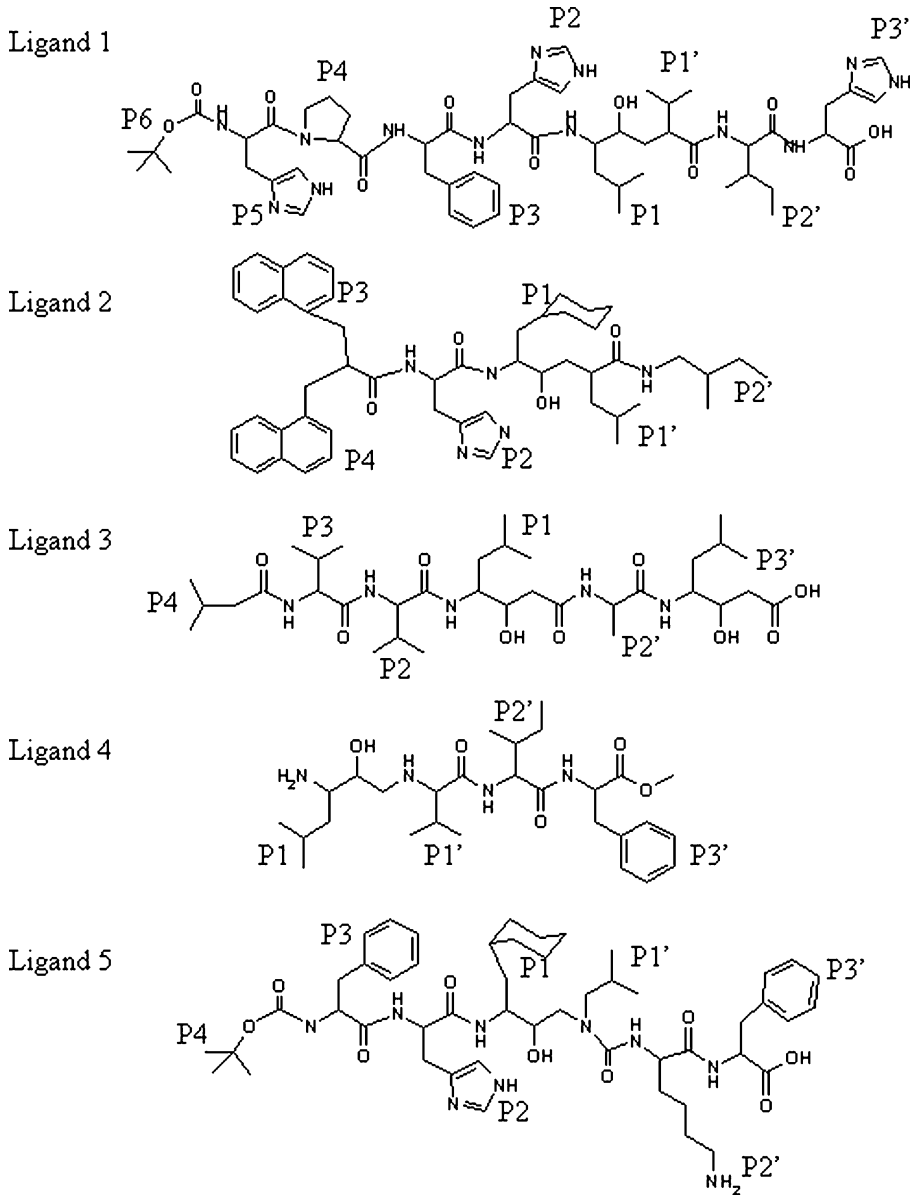

Figure 2. 2D representations of the molecular structure of the five endothiapepsin ligands: (1) H-261, (2) PD-125967, (3) Pepstatin A, (4) BW624, and (5) CP-69799, with labelling of the known interaction sites.

text. Based on the C, N, and O hybridization states, H atoms were added to these five structures using the program Visualizer from Materials Studio [71]. Caution had to be taken for ligand CP-69799 of complex *5er2* (molecule *5*) whose disordered HIS3 sidechain coordinates (located at the level of site P_2 in Figure 2) were not considered at all by the software. To generate various conformations from native structures, we first minimized the energy using the program Discover from Materials Studio [71] with the PCFF force-field, and then carried out a molecular dynamics (MD) run. Three optimization algorithms were successively applied to minimize the energy of the molecular structures: a steepest descent procedure (convergence criteria = 1000 kcal mol^{-1} $Å^{-1}$), a Fletcher-Reeves conjugated gradient approach (convergence criteria = 10 kcal mol^{-1} $Å^{-1}$), and finally a Newton BFGS method (convergence criteria = 0.1 kcal mol^{-1} $Å^{-1}$). The optimized

structures were then used as starting conformations for the MD runs whose last conformation was stored. MD calculations were carried out in the statistical NVE ensemble, at $T = 310$ K, for a period of 10000 steps. The selected time step was 1 fs. As shown in Figure 3, all native inhibitors adopt an extended β-strand conformation without any intramolecular H-bond whereas optimized

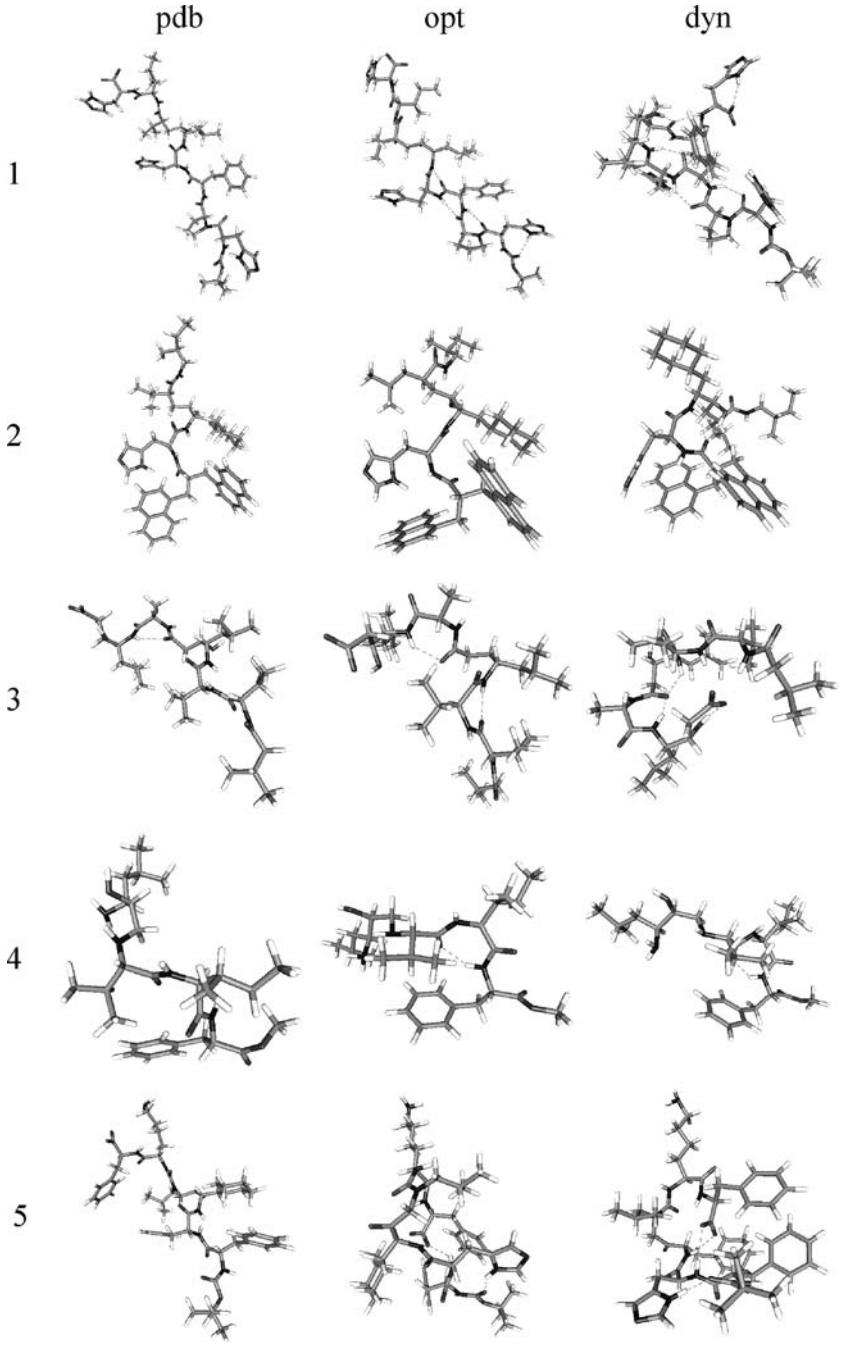

Figure 3. 3D representation of the five endothiapepsin ligands in three different conformations. pdb, opt, and dyn stand for structures in their native, optimized, and dynamic-based conformations. H-bond are displayed using dotted lines.

and MD conformations show additional intramolecular H-bonds with more globular conformations. The effect of temperature is such that dynamic-based structures adopt less stable conformations *vs*. the optimized structures, even if more H-bonds are involved. Additional informations such as the total PCFF energy and the number of intramolecular H-bonds (maximal H-acceptor distance = 2.5 Å, minimal donor-H-acceptor angle = 90°, authorized donors/acceptors = N and O) are given in Table 1. The all-atom root-mean-square distance values, rms_{at}, characterizing each pair of conformations (Table 2) were obtained using the superimposition option implemented in the program gOpenMol [72, 73]. These values will be further used to interpret the quality of the superimpositions of the CP graphs.

Hierarchical clustering results

The hierarchical merging/clustering algorithm that is described in the Theoretical Background section does not require any calculation of the electron density (ED) maps. It is based solely on the knowledge of the analytical expression of the promolecular ED function and its first derivative. The algorithm was applied to the structure of the 15 molecular conformations described above. The decomposition of the structures was achieved at t values ranging from 0 to 10.0 $bohr^2$, with a step of 0.05 $bohr^2$. Each calculation took between 12′ and 39′, on a PC Xeon 32-bit processor.

A complete representation of the merging/clustering results is too heavy to be shown here. However, the dendrogram obtained for molecule *2* in its native conformation, *2_pdb*, is shown in Figure 4. That particular compound was selected due to its reasonable size and the presence of single as well as fused rings. H atoms are not included for clarity, but the merging of H with their chemically bonded atom occurs either at $t = 0.05$ or at $t = 0.1$ $bohr^2$. This is at the origin of the first sudden decrease in the number of local maxima as illustrated in Figure 5. From $t = 0.15$ to 0.50 $bohr^2$, the total number of peaks thus corresponds to the number of non-H atoms in the structures. Then, a second decrease of the number of peaks is observed when the merging occurs between the C and O atoms of the carbonyl functional groups at $t = 0.50$ $bohr^2$. At $t = 0.55$, occurs the merging of the alcohol-O atoms with their bound C atom, which are separated by a larger bond distance than in C=O, e.g., in molecule *2*. At the value of $t = 0.9$ $bohr^2$, the proline five-membered ring of molecule *1* (located at the level of site P_4 in Figure 2) is fully merged and clusters including backbone atoms are

Table 1. Total PCFF energy (kcal/mol) and number of intramolecular H-bonds as calculated using Materials Studio for the 5 considered ligands, in their native (pdb), optimized (opt), and dynamic-based (dyn) conformations.

Ligand	PDB code	pdb		opt		dyn	
		E	No. H-bonds	E	No. H-bonds	E	No. H-bonds
1	2er7	325.7	0	−5.15	5	63.5	6
2	4er1	366.5	0	171.6	0	226.9	2
3	4er2	78.3	1	−95.6	2	−61.9	3
4	5er1	223.7	0	55.6	1	93.7	1
5	5er2	176.4	0	−113.7	3	−44.9	3

Table 2. All-atom root mean square deviations rms_{at} (Å) calculated between pairs of molecular structures using gOpenMol [72, 73].

	Ligand 1	Ligand 2	Ligand 3	Ligand 4	Ligand 5
pdb-opt	3.85	2.39	2.26	2.07	4.12
pdb-dyn	6.64	4.92	4.32	1.72	4.56
opt-dyn	5.27	4.79	3.28	0.74	1.40

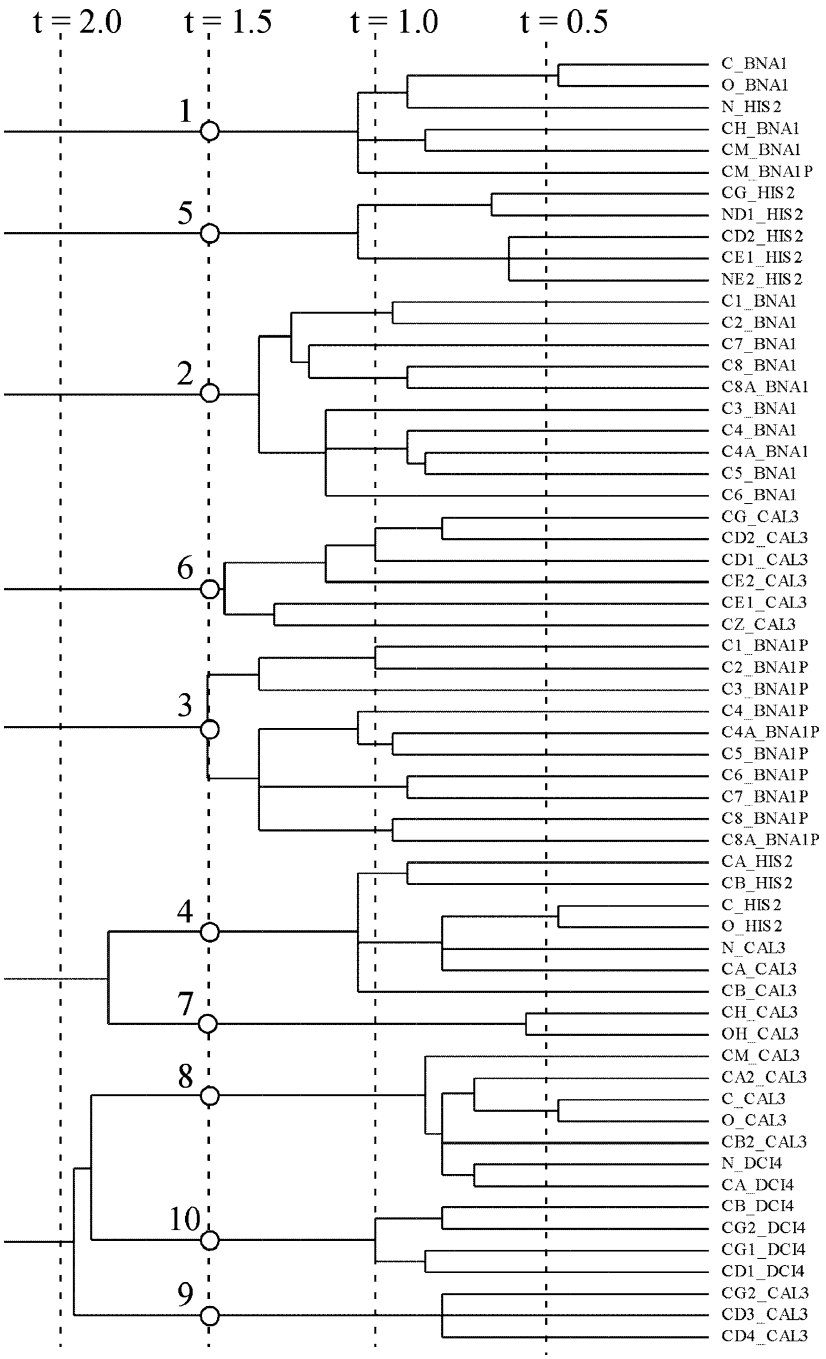

Figure 4. Dendrogram depicting the results of the hierarchical merging/clustering algorithm applied to the PASA electron density distribution of the inhibitor structure *2_pdb* (PD-125967). Results at various values of t (0.5, 1.0, 1.5, and 2.0 bohr^2) are emphasized using vertical dotted lines. Atoms are labeled according to the sequence reported in PDB file *4er1*.

formed in all structures. The merging pattern actually strongly follows what was observed for protein structures [62], i.e., a very interesting situation occurs at $t = 1.5$ bohr^2 where the protein structure is clearly partitioned into backbone and side-chain fragments. At $t = 1.5$ bohr^2, one indeed observes one fragment for each residue backbone, mainly composed of

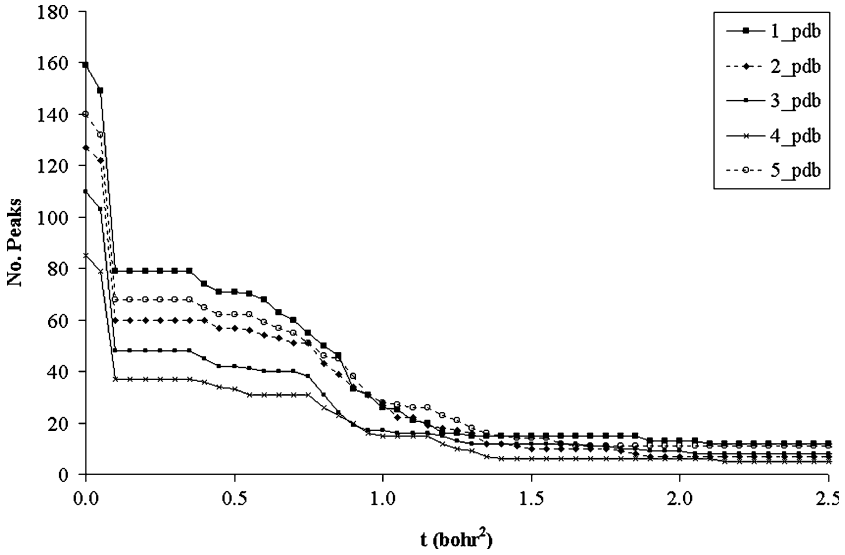

Figure 5. Number of local electron density maxima obtained for the 5 endothiapepsin ligands in their co-crystallized conformation *vs.* the smoothing parameter *t*.

–(C=O)–N–C_α- or a derivative, and one fragment for each residue side-chain. At t = 2.0 bohr^2, some backbone and side-chain fragments are merged, and cyclohexyl rings lead now to a single peak, e.g., in structures *2_opt*, *5_pdb*, and *5_opt*. The merging/clustering patterns observed at t = 1.5 and 2.0 bohr^2 are shown for ligand *2* in its native, optimized, and dynamic-based conformations (Figure 6). 3D representations of the corresponding local maxima obtained for structure *2_pdb* at various t values are shown in Figure 7. Again, this system was selected for display due to its reduced size and variety of chemical groups. For each of the 15 conformations, the number of peaks obtained at various values of t is reported in Table 3.

Figure 6 and Table 3 thus clearly show that, for a given structure, the conformation may affect the number of peaks, even for only slight deformations. This is illustrated, for instance, by structures *4_pdb* and *4_dyn*, whose corresponding rms_{at} value is equal to 0.74 Å only. Also, the number of peaks is not strictly correlated with the number of atoms. For example, in structure *1_pdb*, the 79 local maxima observed at t = 0.2 bohr^2 lead to 26 peaks at t = 1.0, while in structure *2_pdb*, the lower 60 peaks at t = 0.2 bohr^2 lead to a higher number of points than for molecule *1*, i.e., 28, at t = 1.0 bohr^2. In addition, the number of peaks can also vary significantly on a very short range of t values, as for example observed between t = 1.0 and 1.1 bohr^2.

Regarding the peak properties, for a given molecule, ED values are not significantly affected by a conformational change up to t = 1 bohr^2. However, as the smoothing degree increases, the merging pattern which differs with the conformation affects the density values, and this is mostly reflected at t larger than 1 bohr^2. The particular case of molecule *2* is illustrated in Figure 8. One indeed observes, in the three native, optimized, and dynamic-based conformations, density maxima that correspond to similar fragment contents but which do not occur at the same peak number. Also, as illustrated for molecule *2*, a particular chemical group can lead to a single fragment in one conformation (peak #6 in *2_pdb*) but to several fragments in an other conformation (peaks #6, #7, #8 in *2_opt*), and corresponding peaks are thus characterized by very different density values (Figures 6 and 8). In such cases, having a density criterion only in the alignment evaluation function is consequently not sufficient; this justifies the consideration of a geometrical contribution in Equation (7).

In addition to the density ρ, the Laplacian L associated with each PK was also tested as a contribution to the RMS_{AB} function. As for the density, the Laplacian value changes with the fragment content associated with the peaks.

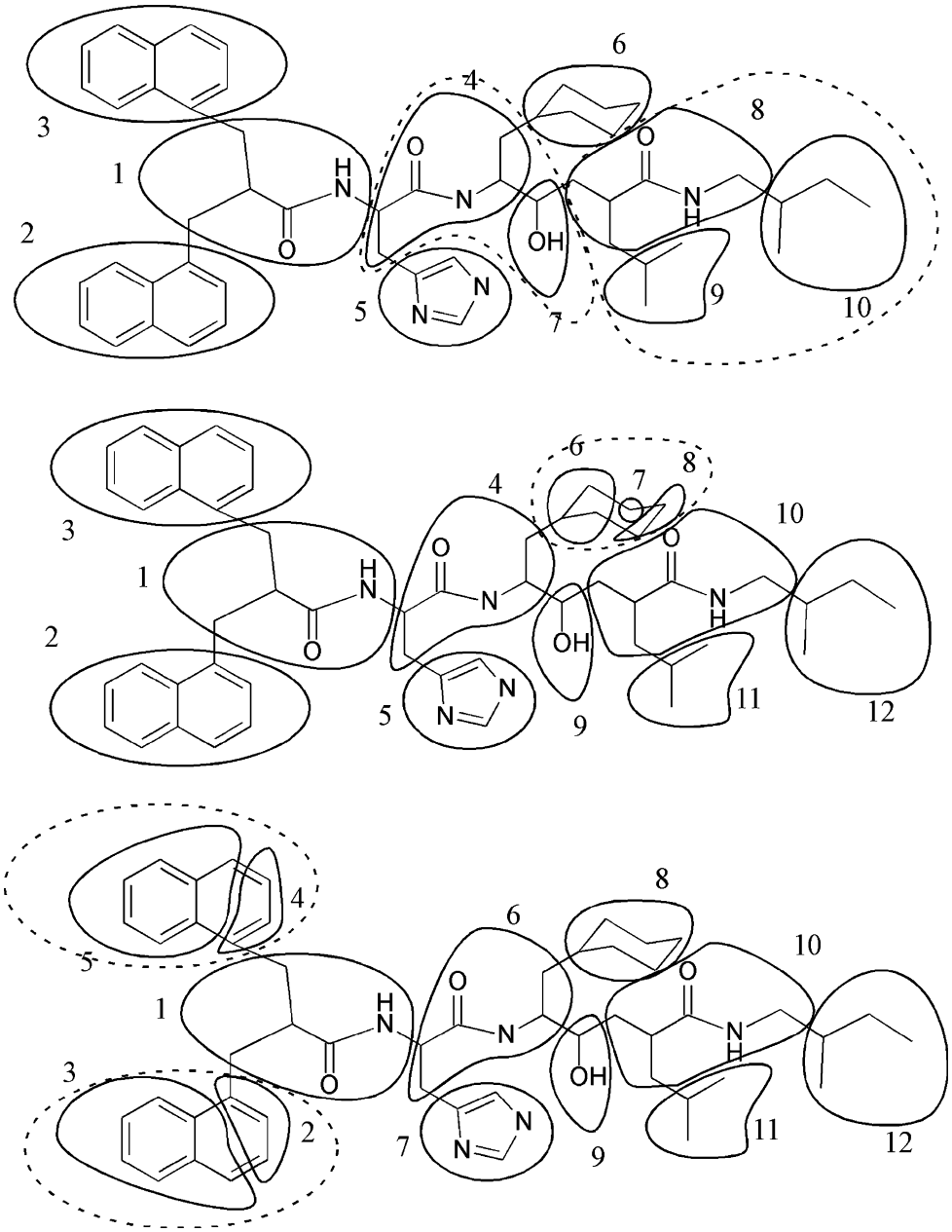

Figure 6. Clusters observed at $t = 1.5$ (plain lines) and 2.0 (dotted lines) bohr^2 for the endothiapepsin ligand 2 (PD-125967) in its PDB native (top), optimized (middle), and dynamic-based (bottom) conformations. Numbers establish the correspondence with the peaks/fragments labelling along the dotted line at $t = 1.5$ bohr^2 in Figure 4.

However, the various alignment tests that were carried out using L (in addition to or in replacement of ρ) did not bring better superposition results than when the ρ and d contributions were used, and the final results that are presented in the two next subsections were all obtained using RMS_{AB} as formulated in Equation (7).

Superposition results of different conformations of a given molecule

In this subsection, we discuss the superposition conditions needed to align graphs of peaks that are characterized by similar density values ρ but different inter-distances. This was achieved to determine

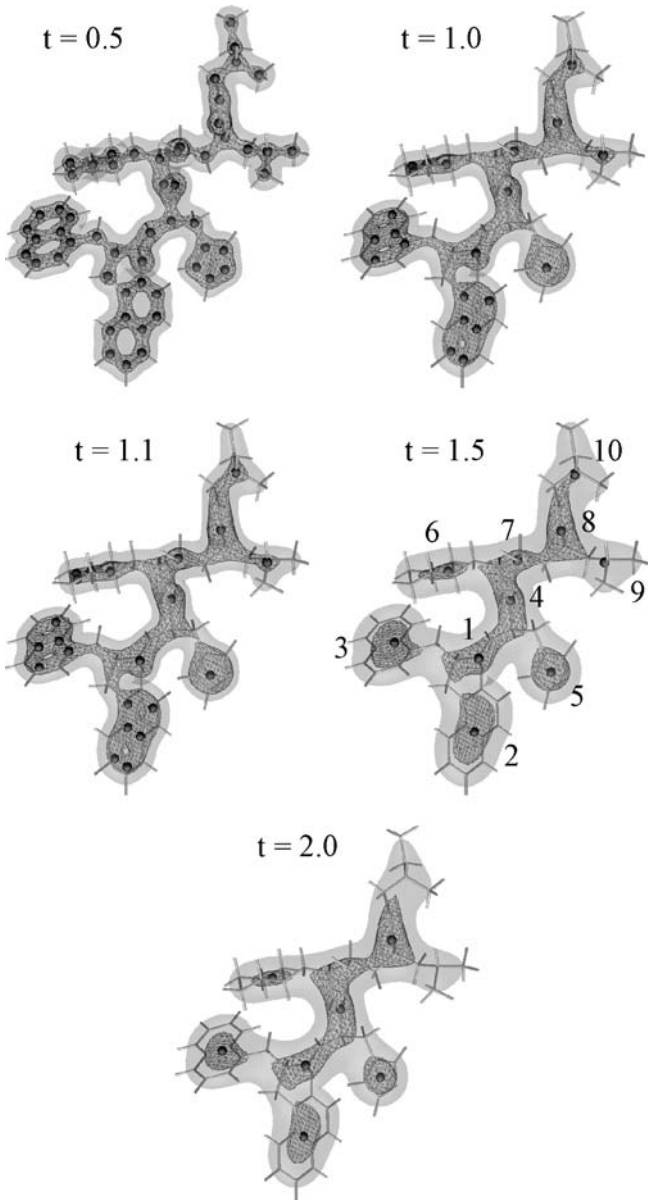

Figure 7. 3D representations of the endothiapepsin ligand structure *2_pdb* (sticks) and corresponding peaks (black spheres) observed in PASA ED distributions smoothed at t = 0.5, 1.0, 1.1, 1.5, and 2.0 bohr^2. Isodensity contours: t = 0.5: 0.2 (triangulated) and 0.1 e^-/bohr^3 (solid); t = 1.0: 0.15 (triangulated) and 0.1 e^-/bohr^3 (solid); t = 1.1: 0.14 (triangulated) and 0.1 e^-/bohr^3 (solid); t = 1.5: 0.125 (triangulated) and 0.075 e^-/bohr^3 (solid); t = 2.0: 0.11 (triangulated) and 0.075 e^-/bohr^3 (solid). Numbers establish the correspondence with peaks/fragments, as in Figures 4 and 6. Figures were obtained using DataExplorer [74].

whether the ρ contribution of Equation (7) allows the recognition of molecular fragments with similar atom contents. Two kinds of inter-distances were tested: Cartesian distances, d, and pseudo-topological distances, d_p, as further detailed.

The first step in evaluating the ability of an ED peak to represent a molecular subgroup was to superpose a molecule onto itself, in a different conformation, and see how the ED peaks were matched. If we consider the subgroups of all ligands as, for example, illustrated in Figure 6 for molecule *2*, best alignment pairs could be determined according to the highest common motif occurring in the matched clusters. For example, at

Table 3. Number of local maxima in PASA ED distributions calculated using a hierarchical merging/clustering algorithm for the five endothiapepsin inhibitors in their native (pdb), optimized (opt), and dynamic-based (dyn) conformations.

t (bohr^2)	Ligand 1			Ligand 2			Ligand 3			Ligand 4			Ligand 5		
	pdb	opt	dyn	pdb	opt	dyn	pdb	opt	dyn	pdb	opt	dyn	pdb	opt	dyn
0.2	79	79	79	60	60	60	48	48	48	37	37	37	68	68	68
0.5	71	71	71	57	57	57	42	42	42	33	33	34	62	61	61
1.0	26	27	29	28	37	30	17	15	14	15	17	16	28	29	31
1.1	21	20	22	22	25	21	16	13	13	15	14	15	26	24	25
1.5	15	16	17	10	12	12	12	12	12	6	7	6	14	14	14
2.0	13	14	14	7	10	10	9	10	10	6	6	6	11	12	11

$t = 1.5$ bohr^2, the best alignment of structures *2_pdb* and *2_opt* involves the following pairs of peaks: (1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (6,7), (6,8), (7,9), (8,10), (9,11), and (10,12). As already mentioned in the previous subsection, point #6 of structure *2_pdb* can be matched to several points of *2_opt*. Indeed, several peaks in one conformation can be associated, based on their atom content, with a single peak in another conformation. We thus also considered the possibility of multiple matches when superposing two molecules.

Based on the best alignments that are expected, we calculated a degree of success for each of the superposition solutions, as:

$$D_s = \frac{n_p}{n_{tot}} \tag{10}$$

where n_p is the number of expected pairs observed among n_{tot}, the total number of matched pairs.

For each ligand, pair superpositions were carried out using the MC/SA algorithm with the input parameters as specified earlier. The final solution with the lowest RMS_{AB} value was kept and the corresponding D_s value was calculated *vs.* the expected solution. In Table 4, we report the number of alignments that are all characterized by $D_s > 0.5$, and $D_s = 1$ for pairs of molecules characterized by an atomic rms_{at} <4 Å (8 alignments), or >4 Å (7 alignments), respectively (See Table 2 for rms_{at} values). The results are presented for various superposition conditions which are combinations of different criteria: w/o standardization ("Stand." or "No-st." in Table 4), w/o

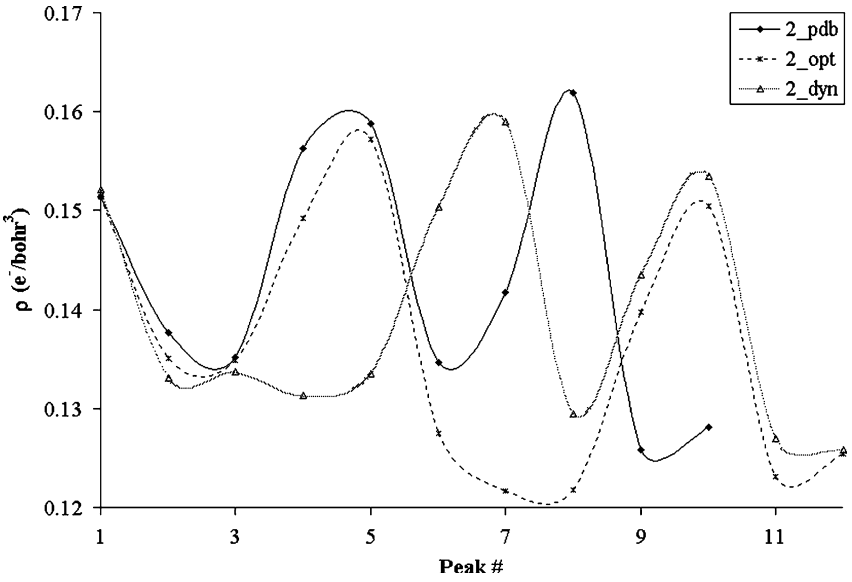

Figure 8. Density values of the peaks observed in smoothed (t = 1.5 bohr^2) PASA ED distributions of molecule 2 in three different conformations: native (pdb), optimised (opt), and dynamic-based (dyn). Lines are displayed as a visualization help.

Table 4. Number of alignments characterized all by $D_s > 0.5$ (a), or $D_s = 1$ (b) obtained using a MC/SA approach for the pair superposition of the five endothiapepsin ligands. Values given in parentheses in each second line correspond to results obtained using d_p, the pseudo-topological distance, instead of d, for the calculation of RMS_{AB} in Equation 7.

	rms_{at} <4 Å								rms_{at} >4 Å							
Total number of alignments:	8								7							
t ($bohr^2$)	1.0		1.1		1.5		2.0		1.0		1.1		1.5		2.0	
Superposition criteria	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
No-st. No-mult. 0-out	5	0	7	1	8	4	8	7	0	0	0	0	1	0	2	0
					(8	5)	(7	6)					(6	2)	(6	3)
No-st. No-mult. 1-out	5	1	6	1	8	7	7	7	1	0	0	0	3	0	2	0
					(8	5)	(7	6)					(6	2)	(7	5)
No-st. No-mult. 2-out	8	1	7	0	7	5	7	5	1	0	0	0	3	0	1	0
					(6	4)	(5	3)					(4	2)	(5	5)
No-st. No-mult. 3-out	8	0	7	0	8	5	7	5	2	0	0	0	3	0	3	0
					(6	4)	(4	2)					(7	2)	(5	5)
No-st. Mult. 0-out	6	0	6	0	6	5	8	5	0	0	0	0	2	0	2	0
					(7	4)	(7	6)					(6	1)	(6	3)
No-st. Mult. 1-out	6	0	5	0	6	5	8	5	0	0	0	0	1	0	0	0
					(8	5)	(7	6)					(4	0)	(7	5)
No-st. Mult. 2-out	7	0	8	0	6	5	5	5	1	0	0	0	1	0	1	0
					(6	3)	(5	3)					(4	1)	(5	5)
No-st. Mult. 3-out	8	0	7	1	8	5	7	4	1	0	0	0	2	0	3	0
					(5	4)	(4	3)					(6	3)	(5	5)
Stand. No-mult. 0-out	8	0	7	2	7	5	6	5	4	0	1	0	3	2	4	2
					(7	7)	(6	6)					(7	3)	(6	2)
Stand. No-mult. 1-out	8	1	8	3	8	7	7	6	4	1	4	1	6	1	6	2
					(7	7)	(6	5)					(7	4)	(5	3)
Stand. No-mult. 2-out	8	2	8	3	6	4	6	4	3	1	4	1	6	2	6	1
					(7	5)	(7	6)					(7	5)	(6	3)
Stand. No-mult. 3-out	8	2	8	3	6	4	5	3	5	1	5	1	6	0	6	1
					(7	5)	(6	5)					(7	4)	(6	4)
Stand. Mult. 0-out	8	0	7	2	7	5	6	3	1	0	2	0	6	0	3	2
					(7	5)	(6	4)					(7	1)	(5	1)
Stand. Mult. 1-out	8	0	8	3	8	6	6	5	3	0	4	0	7	1	6	2
					(8	6)	(6	4)					(7	2)	(5	2)
Stand. Mult. 2-out	7	1	8	3	6	4	6	4	2	0	4	0	4	2	6	1
					(7	4)	(7	6)					(7	2)	(6	3)
Stand. Mult. 3-out	7	1	7	3	6	3	4	2	1	1	3	0	5	1	4	1
					(7	4)	(6	5)					(7	3)	(5	4)

multiple matches (“Mult.” or “No-mult.” in Table 4), and with the “leave-n_{out}-out” technique, where n_{out} varies from 0 to 3 (“n_{out}-out” in Table 4). For example, the notation “No-st. No-mult. 0-out” in the first column of Table 4 describes a calculation that was carried out without standardization, without multiple matches, and considering all points of the smallest graph. The calculation time varies greatly with the number of pairs to match; it evolves between 6″ (4 pairs in molecule *4* under the “leave-two-out” condition when $t = 2.0$ $bohr^2$) to about 100″ (28 pairs in molecule *2* under “the leave-zero-out” condition when $t = 1.0$ $bohr^2$). We recall that the number of CP pairs to be matched depends upon the number of peaks in the smallest CP graph (Table 3).

The analysis of Table 4 leads to the following general conclusions. First, molecular alignments

that correspond to pairs of molecules characterized by rms_{at} <4 Å (Table 2), are closer to the expected results. There are, depending upon the superposition conditions, up to 8 alignments with D_s higher than 0.5. The highest number of fully successful alignments, i.e., alignments with $D_s = 1$, occurs at smoothing degrees of 1.5 and 2.0 bohr^2, with a slightly larger occurrence at $t = 1.5$ bohr^2. More particularly, the detailed data of Figure 9a show that, when rms_{at} <4 Å, all pairs are superposed with at least 50% of success, regardless of the ED smoothing degree, if the ''no-standardization'', ''no-multiple-matches'', and ''leave-two-out'' conditions are applied. At rms_{at} >4 Å, none of the superposition results reaches the value of 100% of success, and the majority is well under the limit of 50%. When focusing more particularly on the results at $t = 1.5$ bohr^2, one can see that all alignments but one lie beyond the limit of 50%, regardless of the rms_{at} value, when the three conditions ''standardization'', ''no-multiple-matches'', and ''leave-two-out'' are used (Figure 9b). However, at $t = 1.5$ bohr^2, the ''no-standardization'', ''no-multiple-matches'', and ''leave-one-out'' approach leads to alignments that are all 80 % successful, when rms_{at} <4 Å (Figure 9c).

On the whole, standardization leads to better results because it implies a rescaling of the ρ contribution to RMS_{AB} *vs.* the d contribution. Indeed, as shown in Table 5, Cartesian distances d have higher means and deviation values than ρ; they are also logically more affected when changing the molecular conformation. The superposition of a given molecule in different conformations, i.e., the recognition of a molecule by itself, is thus more successful when the d contribution is reduced *vs.* the ρ contribution. It is indeed seen, in Table 4, that when standardization is applied (see calculation runs labeled ''Stand.''), the number of good alignments is increased, especially when $rms_{at} > 4$ Å. For example, at $t = 1.5$ bohr^2, all 7 alignments are characterized by $D_s > 0.5$ in run ''Stand. Mult. 1-out'', while there is only 1 in ''No-st. Mult. 1-out''. Therefore, from now on, the data obtained at $t = 1.5$ and 2.0 bohr^2 will be considered for further investigations.

Since the superposition of density CP graphs of a given molecule in different conformations is largely dependent on the way the geometry is represented, it was further decided to test a new d contribution in Equation (7). Rather than calculating Cartesian distances between pairs of peaks, a pseudo-topological distance d_p was evaluated. That new distance is a summation over the successive PK–PK distances d between the adjacent peaks that compose the topological path. Using such a description, d_p between two adjacent peaks is equal to d. This greatly affects the mean distance values and their deviation as illustrated in Table 5. For example, in the case of structure *1_pdb* described at $t = 1.5$ bohr^2, $d_p = 3.33 \pm 12.66$ Å in contrast with $d = 10.24 \pm 5.51$ Å. The alignment results obtained at $t = 1.5$ and 2.0 bohr^2 are strongly improved, especially for the most dissimilar conformations, i.e., when $rms_{at} > 4$ Å (Table 4) whatever the superposition conditions are; on the whole, the number of alignments characterized by $D_s > 0.5$ or $D_s = 1$ is larger when d_p is used than when d is used.

As mentioned previously, tests were also achieved using S_{AB} but they did not bring any improvements with respect to the calculations carried out using RMS_{AB}.

Superposition results of different molecules

The logical next step of the present work is similar to the approach we had already explored before [15]. It consists in the superposition of the ligands in their co-crystallized conformation.

In this subsection, we discuss the superposition conditions needed to align graphs of peaks that have dissimilar density values but close interdistances. The ρ and d contributions are thus likely to affect differently the RMS_{AB} values given in Equation (7) than in the previous subsection.

Reference pair superposition solutions were generated to evaluate the degree of success of the final best alignments provided by the MC/SA superposition approach. These solutions were prepared by properly associating the peaks of the various P_i and $P_{i'}$ sites of the ligands and by visual inspection of the fragment content. The solutions are detailed in Table 6. For each compound, the integers reported in Table 6 correspond to the fragments obtained by our hierarchical merging/clustering algorithm applied to promolecular ED distribution smoothed at $t = 1.5$ bohr^2 (Figures 6 and 10). For example, matched pairs of CPs that participate in the expected alignment of *1_pdb* and *2_pdb* are: (5,1), (6,1), (4,2), (7,3), (8,4), (9,5),

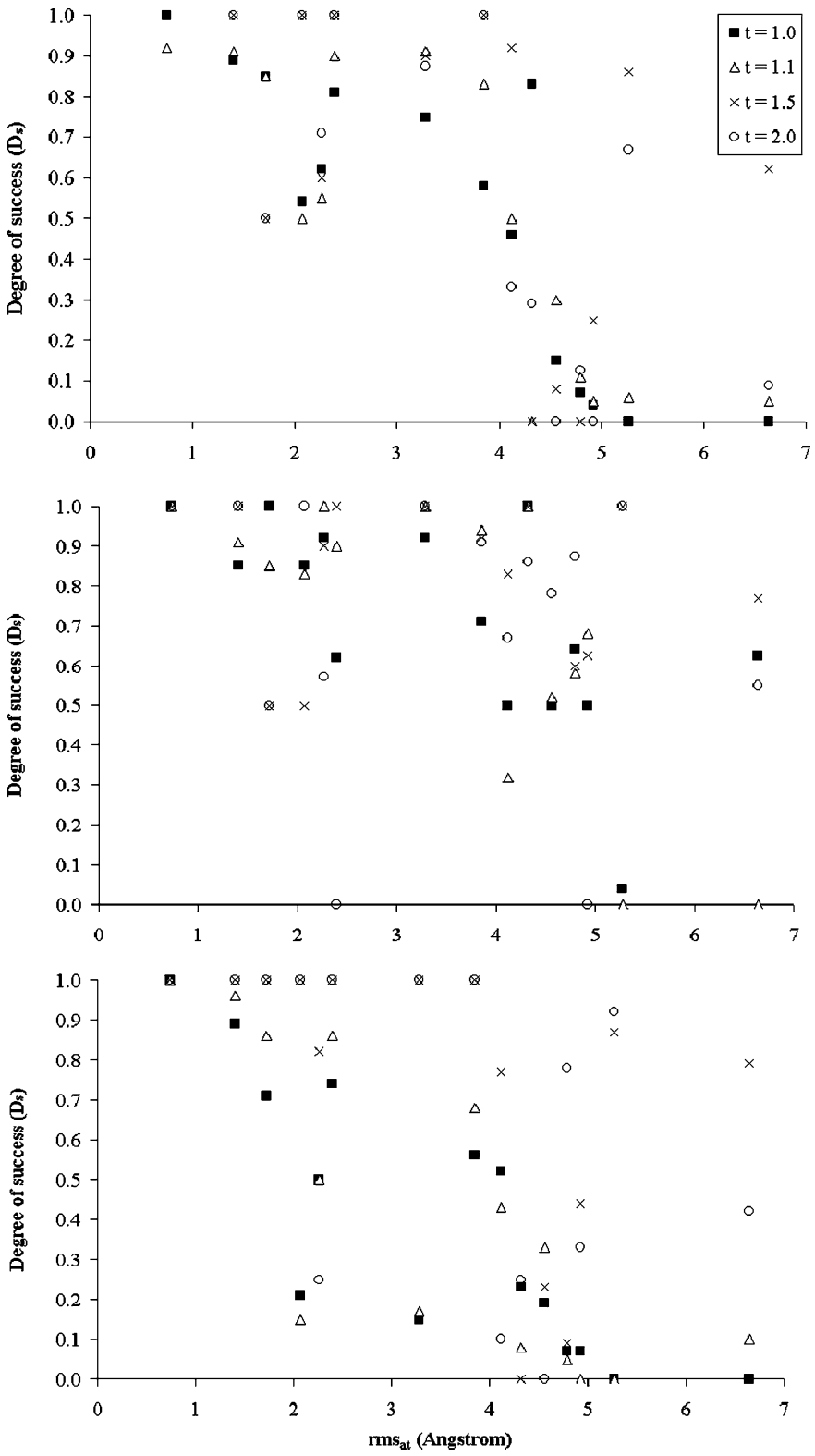

Figure 9. Degree of success of each auto-alignment solution obtained using a MC/SA procedure with the criteria (top) "no-standardization", "no-multiple-matches", and "leave-two-out"; (middle) "standardization", "no-multiple-matches", and "leave-two-out"; and (bottom) "no-standardization", "no-multiple-matches", and "leave-one-out", at various values of *t*.

Table 5. Density (ρ), PK-PK distance (d), and pseudo-topological distance (d_p) means and deviations calculated for the local maxima (PK) obtained for the five endothiapepsin ligands in their native (pdb), optimized (opt), and dynamic-based (dyn) conformations, at t = 1.5 and 2.0 bohr^2.

	ρ (e^-/bohr^3)			d (Å)			d_p (Å)		
	pdb	opt	dyn	pdb	opt	dyn	pdb	opt	dyn
t = 1.5									
1	0.148 ± 0.009	0.148 ± 0.010	0.146 ± 0.012	10.24 ± 5.51	10.40 ± 5.74	7.57 ± 3.09	3.33 ± 12.66	3.90 ± 12.52	3.91 ± 12.75
2	0.143 ± 0.012	0.136 ± 0.012	0.139 ± 0.011	7.53 ± 3.17	7.27 ± 3.05	6.65 ± 2.37	5.42 ± 6.33	6.42 ± 6.16	5.04 ± 6.58
3	0.142 ± 0.011	0.141 ± 0.011	0.141 ± 0.010	7.71 ± 3.88	7.26 ± 3.17	6.25 ± 2.40	2.65 ± 11.38	2.44 ± 11.35	2.90 ± 10.90
4	0.143 ± 0.011	0.140 ± 0.010	0.141 ± 0.011	6.68 ± 2.89	6.35 ± 2.88	6.95 ± 3.12	4.51 ± 5.11	4.58 ± 5.66	4.73 ± 5.61
5	0.141 ± 0.014	0.141 ± 0.014	0.141 ± 0.014	8.40 ± 3.81	6.96 ± 2.71	7.12 ± 2.74	4.13 ± 9.72	3.35 ± 10.93	1.23 ± 13.99
t = 2.0									
1	0.124 ± 0.007	0.123 ± 0.008	0.124 ± 0.008	10.34 ± 5.52	10.50 ± 5.71	7.75 ± 3.06	3.02 ± 13.10	3.69 ± 12.23	3.71 ± 12.10
2	0.125 ± 0.007	0.117 ± 0.008	0.119 ± 0.008	6.90 ± 2.40	7.42 ± 3.04	6.70 ± 2.38	6.93 ± 3.77	6.08 ± 5.94	4.76 ± 6.15
3	0.123 ± 0.002	0.119 ± 0.008	0.119 ± 0.008	7.21 ± 3.80	6.60 ± 2.84	5.76 ± 2.14	2.81 ± 8.36	2.26 ± 8.77	3.99 ± 7.47
4	0.121 ± 0.010	0.118 ± 0.008	0.118 ± 0.009	6.62 ± 2.76	6.42 ± 2.64	6.83 ± 3.03	4.88 ± 4.87	4.94 ± 5.19	5.03 ± 5.27
5	0.121 ± 0.010	0.120 ± 0.010	0.120 ± 0.010	9.05 ± 4.12	6.90 ± 2.73	7.25 ± 2.82	4.35 ± 9.78	3.57 ± 10.57	4.00 ± 10.86

(10,6), (11,7), (12,8), (12,9), and (12,10). In order to check these matches, we generated the corresponding atomic level superpositions through the use of QUATFIT (Figure 11). All these results are similar to the crystalline data and thus show that the reference solutions are valid. The various superposition conditions tested on different conformations were also applied to the alignment of

Table 6. Expected pair alignments between the five endothiapepsin ligands in their native (PDB) conformation, represented as CP graphs of smoothed promolecular ED distributions at t = 1.5 bohr^2. Integers correspond to fragments/peaks as displayed in Figures 6 and 10. For each alignment, the reference molecule is in italic.

1-pdb	5,6	4	7	8	9	10	11	12			
2-pdb	1	2	3	4	5	6	7	8–10			
1-pdb	5	4	6	8,9	8	10	11	12	13	15	14
3-pdb	1	2	3	4	5	6	7	8	9	10	11,12
1-pdb	10	11	12	13	14	15					
4-pdb	1	2	3	4	5	6					
1-pdb	4,5	6	7	8	9	10	11	12	13	14	15
5-pdb	1	2	3	4	5	6,7	8	9,10	11	13	14
3-pdb	1	2	4	6	7	8	9				
2-pdb	1,3	2	4	6	7	8	10				
2-pdb	6	7	8,9	10							
4-pdb	1	2	3	4							
5-pdb	2	1	3	4	5	6,7	8	9	10	12	
2-pdb	1	2	3	4	5	6	7	8	9	10	
3-pdb	6	5,7	8	9	12	10					
4-pdb	1	2	3	4	5	6					
5-pdb	1,3	1	2	4	6,7	8	9	11	14	13	
3-pdb	1	2	3	4,5	6	7	8	9	10	12	
5-pdb	6,7	4,8	9,10	11	12,13	14					
4-pdb	1	2	3	4	5	6					

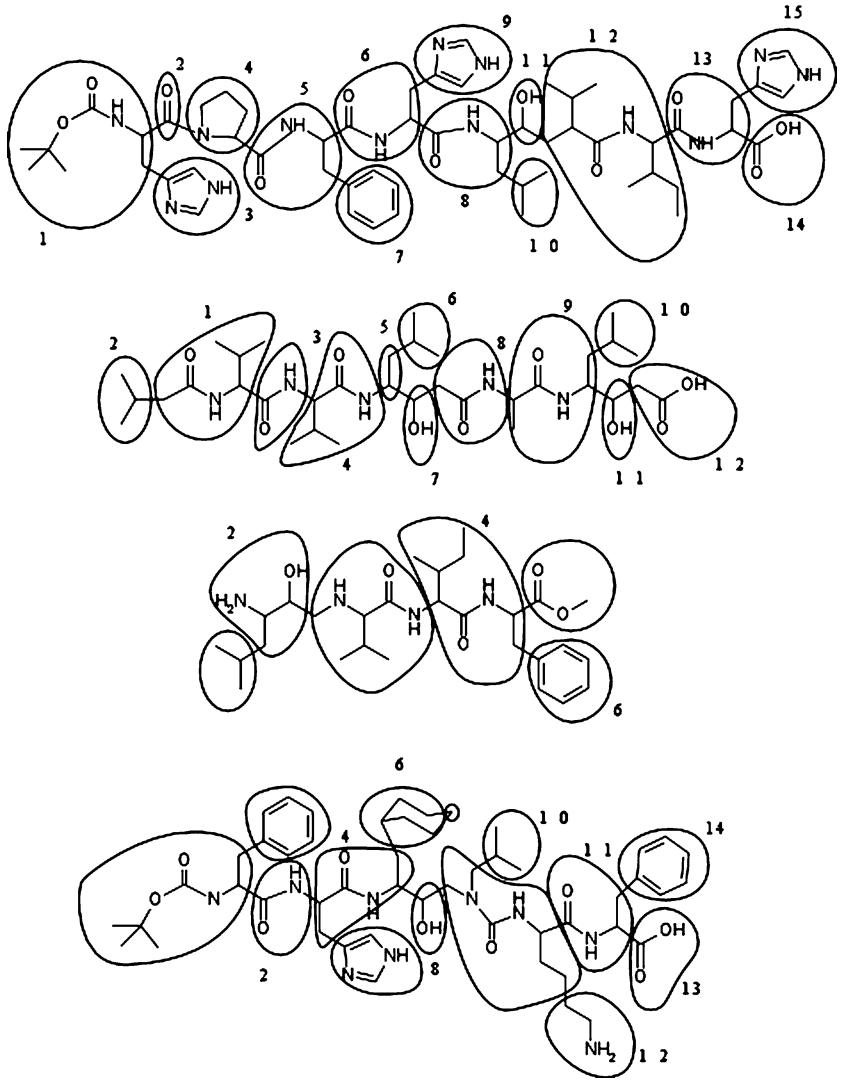

Figure 10. Clusters observed at $t = 1.5$ bohr^2 for the endothiapepsin ligands 1 (H-261), 3 (Pepstatin A), 4 (BW624), and 5 (CP-69799) in their native (PDB) conformations. Ligand 2 (PD-125967) is displayed in Figure 6.

different molecules in their native conformation, i.e., structures named *1_pdb*, *2_pdb*, ... In that particular case, native structures adopt a similar geometry which fits in the receptor active site, while density values may be affected by the chemical composition of the molecules. It is thus at least as important to consider geometry as density in the objective function RMS_{AB}. During the superpositions, two sets of γ limits were used (Equation 6). It was indeed observed that, at the particular value of $t = 1.5$ bohr^2, better solutions could be obtained with γ varying between 0.005 and 0.1, whatever standardization is applied or not. Calculation times varied between 10'' and 31'' depending upon the number of CPs to consider (the lower and largest values of n were equal to 6 and 15, respectively). The numbers of good alignments, i.e., alignments that are characterized by $D_s > 0.5$ or $D_s = 1$, generated under the various superposition criteria are given in Table 7. As directly seen, the number of good alignments is the highest at $t = 1.5$ and 2.0 bohr^2, when the criteria ''no-standardization'' was used. There are indeed up to 7 alignments over a total of 10 that have a value $D_s > 0.5$ (with the criteria ''No-st. Mult. 0_out''), whereas with standardization, the maxi-

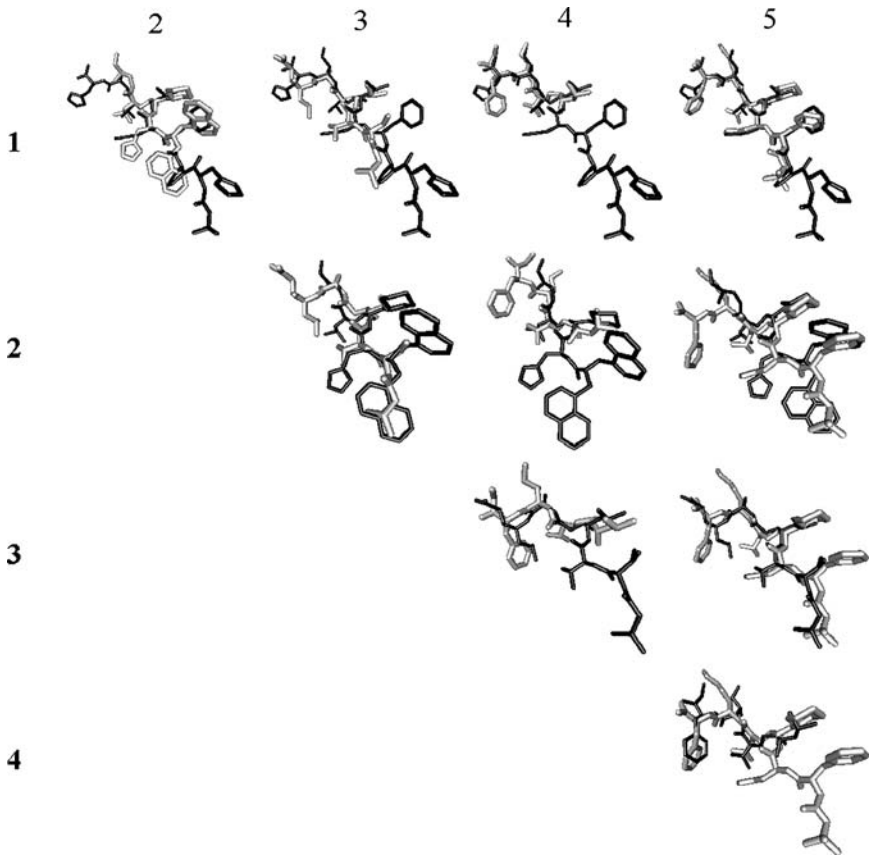

Figure 11. 3D representations of the expected best pair alignments obtained by associating ED peaks at $t = 1.5$ bohr^2 on the basis of the known interaction sites of the five endothiapepsin ligands in their native (PDB) conformation. Molecules numbered along the vertical direction are represented in black.

mum number of alignments with $D_s > 0.5$ is only 3. The absence of any standardization forces the distances to contribute largely to RMS_{AB}. One could even think that, due to the differences in magnitudes (Table 5), the ρ contribution does not play any role. Tests were thus achieved by setting $w_\rho = 0$ in Equation (7); the results showed that even if ρ values are low with respect to d values, they may slightly affect the final superposition results (Table 7). Visual 3D representations corresponding to the best pair alignments obtained at $t = 1.5$ bohr^2 using the criteria ''no-standardization'', ''multiple matches'', and ''leave-zero-out'', generated using QUATFIT [68], are displayed in Figure 12 together with their corresponding RMS_{AB} and D_s values. In addition, we also report the root mean square deviation, rmsd (in Å) obtained between the experimental and computed orientations of the shortest ligand, as well as, in parentheses, the corresponding value reported by Lemmen et al. [45]. We note, at this stage, that our solutions are always the best ones that were obtained, i.e., solutions with a rank equal to 1, which is not necessarily the case in the work by Lemmen and coworkers. With respect to the expected solutions (Figure 11), there are two incorrect alignments ($D_s = 0$), all involving structure *4_pdb*. These alignments however present RMS_{AB} values that are rather small (0.58 and 0.89) when compared to other pair superpositions. As already mentioned in the section Superposition Algorithm, the objective function RMS_{AB} is not strictly correlated with a similarity measure. Let us note that Lemmen et al. reported results for 7 alignments of crystalline structures (1/2, 1/3, 1/4, 1/5, 2/5, 3/5, and 4/5) over a total of 10 [45]. They also met problems in superposing structure *4_pdb* due to its small size. They observed that *4_pdb* is fitted with a reverse orientation onto *5_pdb*, while, in our case, the success is complete. With our

Table 7. Number of alignments characterized all by $D_s > 0.5$ (a), or $D_s = 1$ (b), obtained using a MC/SA approach for the pair superposition of the five endothiapepsin ligands. The alignments were generated using $w_\rho = w_d = 1.0$ in Equation 7, or $w_\rho = 0$ (in parentheses).

Total number of alignments	10							
t (bohr^2)	1.0		1.1		1.5		2.0	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
Superposition criteria								
No-st. No-mult. 0-out	0		1	0	4	1	3	2
			(0	0)	(4	1)	(3	2)
No-st. No-mult. 1-out	0		0	0	4	0	6	3
			(1	0)	(5	1)	(6	3)
No-st. No-mult. 2-out	1	0	1	0	5	0	6	4
	(0	0)	(0	0)	(5	1)	(6	4)
No-st. No-mult. 3-out	0		0	0	6	2	4	2
			(1	0)	(4	2)	(4	2)
No-st. Mult. 0-out	0		0	0	7	1	5	2
			(2	0)	(7	1)	(5	2)
No-st. Mult. 1-out	0		2	0	7	1	7	2
			(2	0)	(7	1)	(7	2)
No-st. Mult. 2-out	0		0		7	2	4	3
					(6	1)	(4	4)
No-st. Mult. 3-out	0	0	0	0	6	2	3	2
	(1	0)	(1	0)	(5	2)	(3	2)
Stand. No-mult. 0-out	1	0	1	0	1	0	1	0
	(0	0)	(0	0)	(0	0)	(0	0)
Stand. No-mult. 1-out	1	0	2	0	2	0	2	0
	(0	0)	(0	0)	(2	0)	(1	1)
Stand. No-mult. 2-out	1	0	0	0	3	0	2	0
	(1	0)	(1	0)	(2	0)	(1	0)
Stand. No-mult. 3-out	1	0	1	0	3	1	1	0
	(0	0)	(1	0)	(3	0)	(1	0)
Stand. Mult. 0-out	1	0	0	0	1	0	1	0
	(0	0)	(1	0)	(1	0)	(2	0)
Stand. Mult. 1-out	1	0	1	0	1	0	2	0
	(0	0)	(0	0)	(3	0)	(1	0)
Stand. Mult. 2-out	0		0	0	3	0	2	0
			(1	0)	(1	0)	(1	0)
Stand. Mult. 3-out	1	0	0	0	3	0	1	0
	(1	0)	(1	0)	(3	0)	(1	0)

approach, the superposition of *4_pdb* on *1_pdb* and *2_pdb* clearly fails. It is however visually successful with *3_pdb* even if the alignment is characterized by $D_s = 0.5$ only. A closer look at the superimposition of the experimental and computed orientations of *4_pdb vs. 3_pdb* shows that the computed orientation is partly displaced with respect to the experimental one (Figure 13).

Finally, let us mention that the use of the pseudo-topological distance in the evaluation of solutions is not efficient. The structures to be compared are all in a similar extended conformations and the best descriptor to be used is thus the Cartesian distance between the ED peaks.

Conclusions

A method for the hierarchical decomposition of a molecular structure based on the analytical knowledge of a promolecular electron density (ED) distribution is developed through the merging/clustering of the ED maxima. It is used to generate graph representations of the molecules at various degrees of details. The locations of the ED peaks at a particular degree of smoothing t are obtained by following the trajectories of the atoms in progressively smoothed ED distributions. The peaks are further linked to generate fully connected graphs.

The analysis of the decomposition patterns was achieved for native, optimized, and dynamic-based conformations of five different endothiapepsin ligands selected after the work of Lemmen et al. [45]. These authors used the superposition program FlexS which involves a scoring function wherein local physicochemical properties (hydrophobicity, partial atomic charges, ...) are distributed using Gaussians centered on atoms or user-defined sites. In the FlexS approach, a reduced number of Gaussians can be obtained, for example, by considering a limited number of functions for the rings, or by merging terminal atoms.

Our results showed that, for a given molecule, the fragment contents and thus the density values of the corresponding ED peaks may vary under a conformational change.

The graph representations so obtained were used in a Monte Carlo/Simulated Annealing procedure for pair superposition. The approach does not require any pre-requisite such as anchor points or closeness to expected solutions. Alignment results are in the form of pairs of matched ED peaks; the visual translation to 3D representations was achieved, when appropriate, using the program QUATFIT [68]. The results showed that the smoothing degrees of $t = 1.5$ and 2.0 bohr^2 led to the most efficient representation levels when combined with an evaluation function which

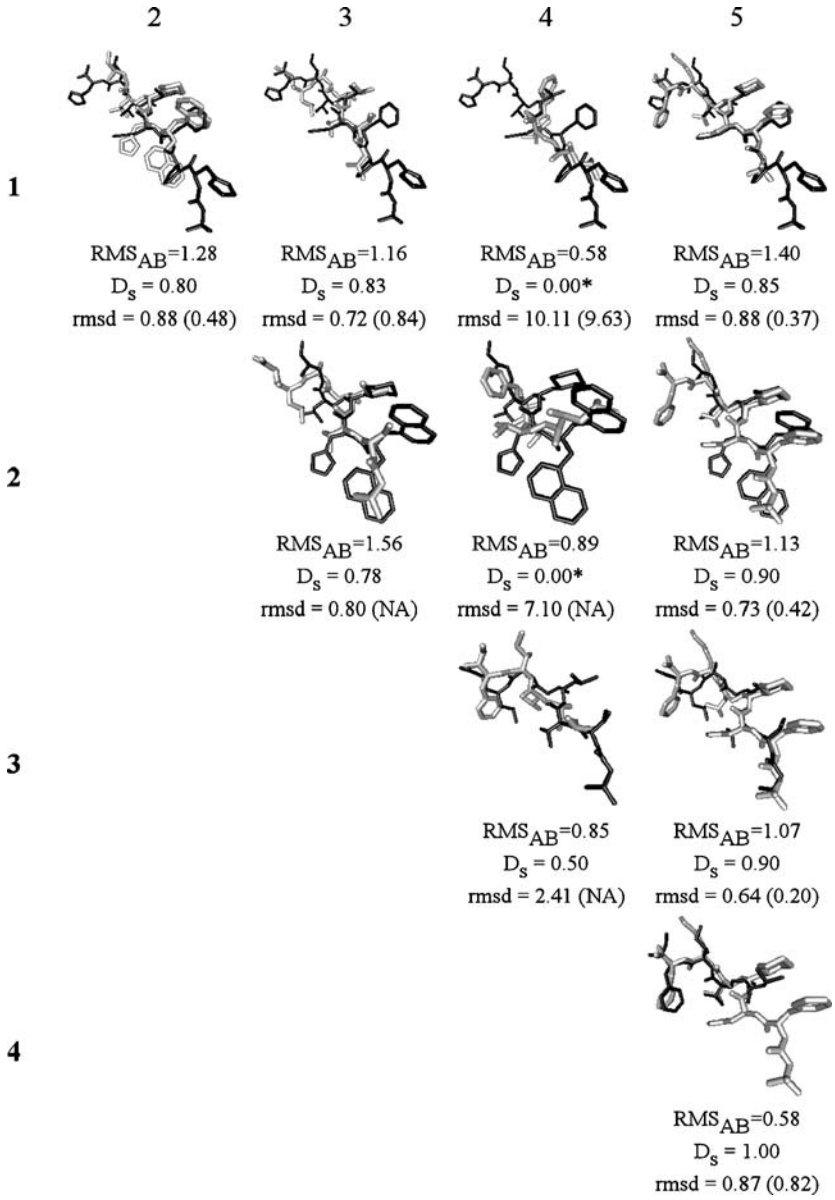

Figure 12. 3D representations of the best pair alignments obtained using a MC/SA procedure with the criteria ''no-standardization, ''multiple-matches'', and ''leave-zero-out'', at the selected value of $t = 1.5$ bohr^2. The two incorrect alignments are marked (*). Molecules numbered along the vertical direction are represented in black. rmsd values are given in Å.

depends on density values and peak inter-distances only. However, it is clear that when two highly different conformations ($rms_{at} > 4$ Å) of a given molecule are aligned, the technique is successful only if the weight of the distance contribution in the evaluation function is reduced with respect to the density contribution, by standardization for example, or by replacing the inter-distance by a pseudo-topological distance along the path defined between two peaks. Besides the change of the geometrical contributions, the density contribution appeared to be well adapted and sufficient to match ED peaks associated with similar chemical contents.

Regarding the superposition of different ligands in their co-crystallized conformations, the best results are obtained with an unmodified evaluation function. The consideration of multiple matches in the alignment procedure is also a useful criterion to generate good alignments.

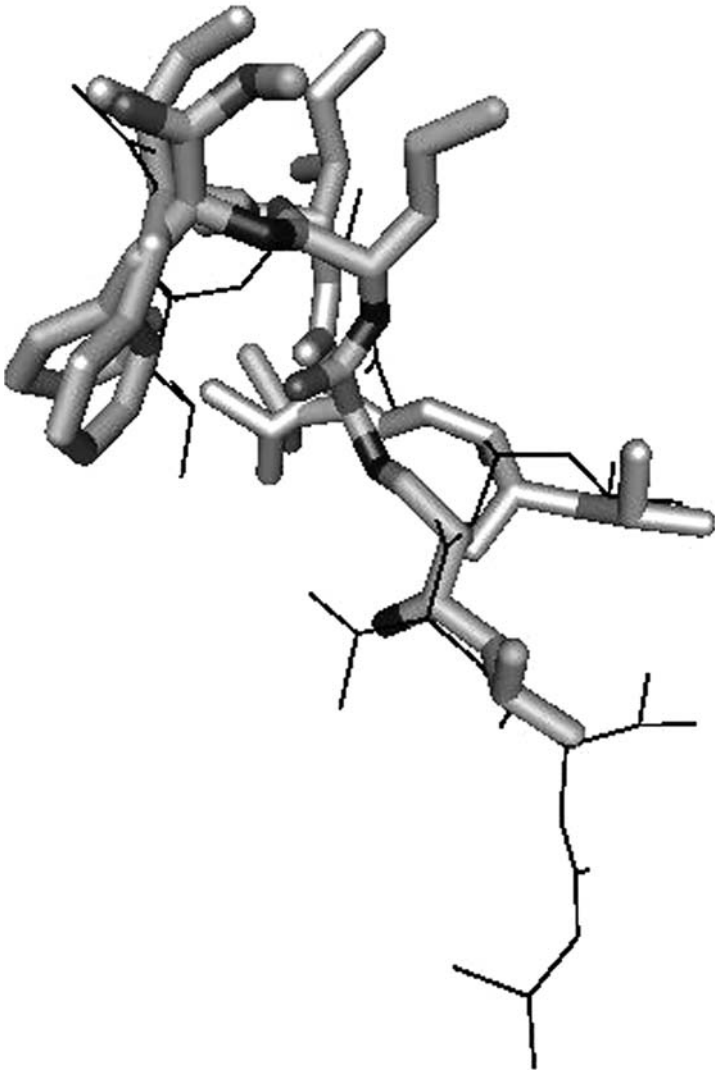

Figure 13. 3D superimposition of the experimental (grey) and computed (atom-based grey shade) orientations of structures *4_pdb vs. 3_pdb* (thin black lines) as obtained using SwissPDBViewer [75] and the approach described in the present paper, respectively.

Acknowledgments

The author thanks Profs S. Fortier and J. Glasgow for continuous interest in their work, Profs. R. Carbó-Dorca, P. Bultinck, and L. Piela for fruitful discussions, as well as Dr. L. Dury for the program DENDRO. The FNRS-FRFC, the ''Loterie Nationale'' (convention n° 2.4578.02), and the FUNDP, are gratefully acknowledged for the use of the Interuniversity Scientific Computing Facility (ISCF) Center. NM thanks the ''Fonds National de la Recherche Scientifique'' for her Scientific Research Worker position.

References

1. Robinson, D.D., Barlow, T.W. and Richards, W.G., J. Chem. Inf. Comput. Sci., 37 (1997) 943.
2. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Comput. Chem., 18 (1997) 934.
3. Carbó, R., Calabuig, B., Vera, L. and Besalú, E., Adv. Quantum Chem., 25 (1994) 253.
4. Mezey P.G., In Johnson M.A. and Maggiora G.M. (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990, pp. 321–368.
5. Bader, R.W. Atoms in Molecules – A Quantum Theory. Clarendon Press, Oxford, 1995.
6. Popelier, P.L.A, J. Phys. Chem. A, 103 (1999) 2883.
7. Gadre, S.R. and Pundlik, S.S., J. Am. Chem. Soc., 117 (1995) 9559.

8. Leboeuf, M., Köster, A.M., Jug, K. and Salahub, D.R., J. Chem. Phys., 111 (1999) 4893.
9. Wild, D.J. and Willett, P., J. Chem. Inf. Comput. Sci., 36 (1996) 159.
10. Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., J. Chem. Inf. Comput. Sci., 36 (1996) 900.
11. Meurice N., Leherte L., Vercauteren D.P., Bourguignon J.-J. and Wermuth C., In van de Waterbeemd H., Testa B. and Folkers G. (Eds.), Computer-Assisted Lead Finding and Optimization, Verlag, Basel, 1997, pp. 497–510.
12. Meurice N., Leherte L. and Vercauteren D.P., In Devillers J. (Ed.), SAR and QSAR in Environmental Research Vol. 8, OPA, Amsterdam, 1998, pp. 195–232.
13. Leherte L., Meurice N. and Vercauteren D.P., In Mastorakis N. (Ed.), Mathematics and Computers in Modern Science. Acoustics and Music, Biology and Chemistry, Business and Economics, World Scientific Engineering Society, Athens, 2000, pp. 158–164.
14. Leherte, L., Meurice, N. and Vercauteren, D.P., J. Chem. Inf. Comput. Sci., 40 (2000) 816.
15. Leherte, L., J. Math. Chem., 29 (2001) 47.
16. Glick, M., Robinson, D.D., Grant, G.H. and Richards, W.G., J. Am. Chem. Soc., 124 (2002) 2337.
17. Glick, M., Grant, G.H. and Richards, W.G., J. Med. Chem., 45 (2002) 4639.
18. Allen, F.H., Acta Cryst. B, 58 (2002) 380.
19. Kramer, B., Rarey, M. and Lengauer, Th., PROTEINS: Struct. Funct. Genet., 37 (1999) 228.
20. Kalászi, A. and Farkas, Ö., J. Mol. Struct. (THEOCHEM), 666–667 (2003) 645.
21. Abrahamian, E., Fox, P.C., Naerum, L., Christensen, I. Th., Thogersen, H. and Clark, R.D., J. Chem. Inf. Comput. Sci., 43 (2003) 458.
22. Klebe, G., Mietzner, Th. and Weber, F., J. Comput. Aided Mol. Des., 8 (1994) 751.
23. Klebe, G., Mietzner, Th. and Weber, F., J. Comput. Aided Mol. Des., 13 (1999) 35.
24. Mestres, J., Rohrer, D.C. and Maggiora, G.M., J. Comput. Aided Mol. Des., 14 (2000) 39.
25. Makino, S. and Kuntz, I.D., J. Comput. Chem., 19 (1998) 1834.
26. Lin, T.-H., Lin, J.-J. and Lu, Y.-J., Biochim. Biophys. Acta, 1429 (1999) 476.
27. Feher, M. and Schmidt, J.M., Chem. Inf. Comput. Sci., 40 (2000) 495.
28. Ghose, A.K. and Crippen, G.M., Comput. Chem., 6 (1985) 350.
29. Clark, D.E., Willett, P. and Kenny, P.W., J. Mol. Graphics, 10 (1992) 194.
30. Clark, D.E., Willett, P. and Kenny, P.W., J. Mol. Graphics, 11 (1993) 146.
31. Wildman, S.A. and Crippen, G.M., J. Mol. Graphics Modell., 21 (2002) 161.
32. Raymond, J.W. and Willett, P., J. Chem. Inf. Comput. Sci., 43 (2003) 908.
33. Mills, J.E.J, de Esch, I.J.P., Perkins, T.D.J. and Dean, P.M., J. Comput. Aided Mol. Des., 15 (2001) 81.
34. Jones, G., Willett, P. and Glen, R.C., J. Comput. Aided Mol. Des., 9 (1995) 532.
35. Handschuh, S., Wagener, M. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 38 (1998) 220.
36. Labute, P., Williams, Ch., Feher, M., Sourial, E. and Schmidt, J.M., J. Med. Chem., 44 (2001) 1483.
37. Chae, C.H., Oh, D.G. and Shin, W., J. Comput. Chem., 22 (2001) 888.
38. Korhonen, S.-P., Tuppurainen, K., Laatikainen, R. and Peräkylä, M., J. Chem. Inf. Comput. Sci., 43 (2003) 1780.
39. Gironés, X. and Carbó-Dorca, R., J. Comput. Chem., 25 (2004) 153.
40. Szabó, Z., Vargyas, M. and Johnson, A.P., J. Chem. Inf. Comput. Sci., 40 (2000) 339.
41. Pitman, M.C., Huber, W.K., Horn, H., Krämer, A., Rice, J.E. and Swope, W.C., J. Comput. Aided Mol. Des., 15 (2001) 587.
42. Jain, A.N., J. Med. Chem., 46 (2003) 499.
43. Krämer, A., Horn, H.W. and Rice, J.E., J. Comput. Aided Mol. Des., 17 (2003) 13.
44. Lemmen, Ch. and Lengauer, Th., J. Comput. Aided Mol. Des., 11 (1997) 357.
45. Lemmen, Ch., Lengauer, Th. and Klebe, G., J. Med. Chem., 41 (1998) 4502.
46. Fradera, X., Knegtel, R.M.A. and Mestres, J., PROTEINS: Struct. Funct. Genet., 40 (2000) 623.
47. Good, A.C. and Richards, W.G., J. Chem. Inf. Comput. Sci., 33 (1993) 112.
48. Tsirelson, V.G., Avilov, A.S., Abramov, Y.A., Belokoneva, E.L., Kitaneh, R. and Feil, D., Acta Cryst. B, 54 (1998) 8.
49. Tsirelson, V., Abramov, Y., Zavodnik, V., Stash, A., Belokoneva, E., Stahn, J., Pietsch, U. and Feil, D., Struct. Chem., 9 (1998) 249.
50. Gironés, X., Amat, L. and Carbó-Dorca, R., J. Mol. Graphics Modell., 16 (1998) 190.
51. Botella, V. and Pacios, L.F., J. Mol. Struct. (THEOCHEM), 426 (1998) 75.
52. Mitchell, A.S. and Spackman, M.A., J. Comput. Chem., 21 (2000) 933.
53. Gironés, X., Carbó-Dorca, R. and Mezey, P.G., J. Mol. Graphics Modell., 19 (2001) 343.
54. Downs, R.T., Gibbs, G.V., Boisen, M.B. Jr. and Rosso, K.M., Phys. Chem. Miner., 29 (2002) 369.
55. Bultinck, P., Carbó-Dorca, R. and Van Alsenoy, Ch., J. Chem. Inf. Comput. Sci., 43 (2003) 1208.
56. Aubert, E., Porcher, F., Souhassou, M. and Lecomte, Cl., Acta Cryst. B, 59 (2003) 687.
57. Amat, L. and Carbó-Dorca, R., J. Comput. Chem., 18 (1997) 2023.
58. Coefficients and exponents can be seen and downloaded from the Web site: http://iqc.udg.es/cat/similarity/ASA/funcset.html.
59. Kostrowicki, J., Piela, L., Cherayil, B.J. and Scheraga, H.A., J. Phys. Chem., 95 (1991) 4113.
60. Johnson, C.K., ORCRIT. The Oak Ridge Critical Point Network Program, Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, TN, 1977.
61. Leherte, L., Dury, L. and Vercauteren, D.P., J. Phys. Chem. A, 107 (2003) 9875.
62. Leherte, L., Acta Cryst. D, 60 (2004) 1254.
63. Leung, Y., Zhang, J.-S. and Xu, Z.-B., IEEE T. Pattern Anal., 22 (2000) 1396.
64. Gilbert, D.G., Phylodendron, for Drawing Phylogenetic Trees, Version 0.8d, Indiana University, 1996. Software at http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/. Web form at http://iubio.bio.indiana.edu/treeapp/treeprintform.html.
65. Dury, L. ENDRO. Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium, 2002.
66. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., J. Chem. Phys., 21 (1953) 1087.

67. Kirkpatrick, S., Gelatt, C.D. Jr. and Vecchi, M.P., Science, 220 (1983) 671.
68. Heisterberg, D.J., Technical report, Ohio Supercomputer Center, Columbus (OH). Translation from FORTRAN to C and Input/Output by Jan Labanowski, Ohio Supercomputer Center, Columbus (OH), 1990.
69. Bailey, D. and Cooper, J.B., Protein Sci., 3 (1994) 2129.
70. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235; http://www.rcsb.org/pdb.
71. http://www.accelrys.com.
72. Laaksonen, L., J. Mol. Graphics, 10 (1992) 33.
73. Bergman, D.L., Laaksonen, L. and Laaksonen, A., J. Mol. Graphics Modell., 15 (1997) 301.
74. Open Visualization Data Explorer, v. 4.2.0, http://www.research.ibm.com/dx.
75. Guex, N. and Peitsch, M.C., Electrophoresis, 18 (1997) 2714; http://www.expasy.org/spdbv.