Theor Chem Acc (2012) 131:1259
DOI 10.1007/s00214-012-1259-y

REGULAR ARTICLE

Smoothed Gaussian molecular fields: an evaluation of molecular alignment problems

Laurence Leherte · Daniel P. Vercauteren

Received: 29 February 2012 / Accepted: 11 July 2012 / Published online: 28 July 2012

Abstract Several smoothed Gaussian-based descriptors used in a molecular superposition algorithm are presented. One descriptor, as detailed in a previous work (Leherte in J Comput Chem 27:1800–1816, 2006), is the full electron density approximated through the promolecular atomic shell approximation (PASA) (Amat and Carbó-Dorca in J Chem Inf Comput Sci 40:1188–1198, 2000). Herein, we additionally present a new descriptor, that is, the charge density of a molecule calculated via the Poisson equation. The Coulomb potential as approximated by Good et al. (J Chem Inf Comput Sci 32:188–191, 1992) and atom-based functions such as hydrogen bond donor or acceptor properties, lipophilicity as detailed in the work of Totrov (Chem Biol Drug Des 71:15–27, 2008) were also considered. A Monte Carlo/Simulated Annealing superposition method is applied to a set of six families of drug molecules, that is, elastase inhibitors, ligands of endothiapepsins, trypsins, thermolysins, p38 MAP kinases, and rhinovirus, all of them already reported in the literature, for discussing superposition problems. The results show that the descriptor selection can be guided by the nature of the interactions expected to occur between the drug molecules and their receptor. They also emphasize the particular efficiency of the PASA descriptor for molecules characterized by significant shape properties.

Keywords Promolecular electron density distribution · Poisson equation · Coulomb potential · Smoothing · Molecular alignment · Similarity index

1 Introduction

Since at least two decades, the use of Gaussian functions for the evaluation of the molecular similarity has been an attractive strategy as both it allows short calculation times and it is very easy to implement [1]. Indeed, using such functions, similarity measures are directly related to distances between the atoms that constitute the molecular structures to be compared [2, 3]. In molecular modelling, it is indeed common to access the shape of a molecule by fitting spheres at the atom locations, such as the well-known van der Waals (vdW) spheres [4], or by considering Gaussian functions as presented, for example, by Good and Richards [5] and Grant and Pickup [6–8]. In order to consider some hardness of the atoms when bound in a molecule, Good and Richards [5] used atomic electron density (ED) distributions, which are set to zero beyond the vdW radius of the atoms. Grant et al. [6–8] compared the conventional hard sphere representation and a Gaussian-based model and proposed applications in the field of shape comparison using the so-called shape multipoles or moments. Their work led, notably, to the implementation of the program ROCS [9]. Later, applications were reported by Haigh et al. [10] who suggested a transferable and fast shape fingerprint approach based on Grant et al.'s works. Maggiora and coworkers [11] used spherically symmetric Gaussian functions located on selected atoms

Published as part of the special collection of articles celebrating theoretical and computational chemistry in Belgium.

Electronic supplementary material The online version of this article (doi:10.1007/s00214-012-1259-y) contains supplementary material, which is available to authorized users.

L. Leherte (✉) · D. P. Vercauteren
Laboratoire de Physico-Chimie Informatique, Unité de Chimie Physique Théorique et Structurale, Facultés Universitaires Notre-Dame de la Paix (FUNDP), Rue de Bruxelles 61, 5000 Namur, Belgium
e-mail: laurence.leherte@fundp.ac.be

and characterized by adjustable magnitudes and widths to modulate the degree of details needed to achieve protein alignments. Within such a representation, each amino acid of a protein is described by a linear combination of a limited number of Gaussians. Their approach was implemented in the program MIMIC [12]. Duncan and Olson [13] also defined molecular surfaces as a sum over atomic Gaussian functions. In their work, emphasis was given on the resolution of these surfaces, which can be modified by convolving the ED distribution function with a Gaussian function of selected variance. On such bases, various applications were proposed in the fields of molecular comparison and molecular complementarity, as well as in visual interpretation of molecular surfaces. Klebe et al. [14] established a mathematical formalism for the evaluation of molecular similarity in three-dimensional (3D) QSAR studies. Similarity was evaluated between a given molecule and a spherical probe and was calculated at each point of a 3D grid as a summation over atomic contributions [14]. More recently, Totrov [15] described a molecule through seven 3D atomic property fields (APFs) calculated from Gaussians functions centred on the constituting atoms. These atomic properties are hydrogen bond donor, hydrogen bond acceptor, sp^2 hybridization, lipophilicity, size, charge, and electronegativity. The author identified 21 atom types and associated them a value for each of those seven properties. Totrov's approach is close to the procedure described by Lemmen et al. [16] in the program FLEXS wherein physicochemical properties such as hydrophobicity, charge, hydrogen bonding are approximated by a set of Gaussian functions centred on atoms or other regions defined by the user. Proschak et al. [17] applied a molecular shape description expressed as a summation over atomic Gaussian functions to define molecular surface elements useful in surface matching calculations and implemented their approach in the program "Shapelets". Chan et al. [18] used Gaussians to define a scoring function for their alignment procedure. The function is calculated as a summation over Gaussian terms depending upon the distances occurring between atoms characterized by a given property (size, hydrophobicity). The authors implemented their scoring function in the program MOE [19].

Superposition of molecules is a problem that involves many local solutions. A way to reduce the number of possible alignments is to lower the resolution of the molecular field under consideration, in order words, to lower the level of details by smoothing the 3D scalar field [11, 13, 20]. This can easily be achieved through a convolution product with a Gaussian function, as proposed by Kostrowicki et al. [21]. Another approach consists in limiting the number of points representing the molecules, such as in the studies of Glick et al. [22, 23], wherein the atoms are clustered based on their separating distances. We also used such an approach by representing molecular systems as graphs of smoothed ED critical points [24, 25].

Following a work we previously achieved on molecular similarity of promolecular ED distribution functions [20], we expand here the concepts presented before through the calculation of charge density (CD) distributions calculated from smoothed electrostatic potential functions via the Poisson equation. The advantage of Gaussian functions in evaluating various similarity measures is again considered to easily calculate integrals such as the overlap and the so-called Laplacian ones at low cost.

In this paper, we treat various molecular similarity problems through the study of six different families of molecules, as already detailed in the literature. The selected families are the TOMI and DFKi elastase ligands [26–29], inhibitors of endothiapepsins [16, 20, 30], trypsins [16, 18, 31, 32], thermolysins [16, 17, 30–32], human rhinovirus HRV14 [16, 18, 32], and of p38 mitogen-activated proteins (MAP) [18, 32].

We first used a promolecular description of the ED distribution function of the various molecules, as reported before [20]. We also apply the formalism obtained for the CD calculated from smoothed electrostatic potential functions through the Poisson equation. Such a CD distribution function was previously considered to design, through its topological properties, reduced point charge models for proteins [33]. This new aspect is considered in comparison with the method described by Good et al. [34] to superpose Coulomb potential functions and implemented by us in combination with a smoothing approach. Finally, we also considered a smoothed version of the APFs developed by Totrov [15].

Flexibility is not considered in the present work as we compare alignments to discuss the efficiency of the various smoothed molecular fields under consideration without the influence of the conformation. As shown by the results, the various molecular fields can provide different results and their efficiency can vary with the nature of the interactions involved between the molecules and their receptor.

In the next section, we briefly recall the mathematical expressions needed to superpose the molecules and to evaluate the corresponding similarity degree. We detail the new expressions related to CD distribution functions and smoothing in general. Thereafter, we present the molecular systems under study and discuss the alignment results. Conclusions and perspectives are provided at the end of the paper.

2 Theoretical background

In this section, it is described how smoothed Gaussian-based scalar fields can be calculated analytically and how similarity measures and indices are evaluated.

2.1 Promolecular electron density distributions

In their work related to the Promolecular Atom Shell Approximation (PASA), Amat and Carbó-Dorca used atomic Gaussian ED functions that were fitted on 6-311G atomic basis set results [35]. In the PASA approach that is considered in the present work, a promolecular ED distribution ρ_A is represented analytically as a weighted summation over the nat atomic ED distributions ρ_a, which are described in terms of series of three squared 1s Gaussian functions fitted from atomic basis set representations [36]:

$$\rho_A = \sum_{a\in A}^{\text{nat}} \rho_a \tag{1}$$

with:

$$\rho_a(\mathbf{r}) = Z_a \sum_{i=1}^{3} w_{a,i} \left[\left(\frac{2\varsigma_{a,i}}{\pi} \right)^{34} \mathrm{e}^{-\varsigma_{a,i}|\mathbf{r}-\mathbf{R}_a|^2} \right]^2 \tag{2}$$

where Z_a, $\mathbf{R}_a$, and $w_{a,i}$ and $\varsigma_{a,i}$, are the atomic number of atom a, its position vector, and the two fitted parameters, respectively.

To generate smoothed 3D ED functions, ρ_A is directly expressed as the solution of the diffusion equation according to the formalism presented by Kostrowicki et al. [21]:

$$\rho_{a,t}(\mathbf{r}) = \sum_{i=1}^{3} s_{a,i} \quad \text{where} \quad s_{a,i} = \alpha_{a,i} \mathrm{e}^{-\beta_{a,i}|\mathbf{r}-\mathbf{R}_a|^2} \tag{3}$$

with:

$$\alpha_{a,i} = Z_a w_{a,i} \left(\frac{2\varsigma_{a,i}}{\pi} \right)^{3/2} \frac{1}{\left(1+8\varsigma_{a,i}t\right)^{3/2}} \quad \text{and} \quad \beta_{a,i} = \frac{2\varsigma_{a,i}}{\left(1+8\varsigma_{a,i}t\right)} \tag{4}$$

where t is the smoothing degree of the ED, unsmoothed EDs being obtained by imposing $t = 0$ bohr^2.

When using the PASA description, only the non-hydrogen atoms of the molecular structures are considered. It is done so to limit the calculation time of the alignment procedures. The advantage of such a descriptor thus relies in the fact that no a priori knowledge of the protonation state of the molecules is required.

2.2 Coulomb potential and charge density distribution functions

The electrostatic potential function generated by a molecule A is approximated by a summation over its atomic contributions using the Coulomb equation:

$$\Phi_A(\mathbf{r}) = \sum_{a\in A}^{\text{nat}} \frac{q_a}{|\mathbf{r}-\mathbf{R}_a|} \tag{5}$$

q_a being the net charge of atom a. A smoothed version of the potential generated by atom a, $\Phi_{a,t}(r)$, can be expressed as [37]:

$$\Phi_{a,t}(r) = \frac{q_a}{r} \operatorname{erf}\left(\frac{r}{2\sqrt{t}} \right) \tag{6}$$

where t is the smoothing parameter and erf stands for the error function. From the potential given in Eq. 6, the corresponding analytical CD function $\rho_{a,t}(r)$ can be obtained from the Poisson equation:

$$-\nabla^2 \Phi_{a,t} = \frac{\rho_{a,t}}{\varepsilon_0} \tag{7}$$

and expressed as:

$$\rho_{a,t}(r) = \frac{q_a}{(4\pi t)^{3/2}} \mathrm{e}^{-r^2/4t} \tag{8}$$

In such a formalism, $\rho_{a,t}(r)$ cannot be calculated at $t = 0$. Indeed, that situation corresponds to the original Coulomb potential for which the solution of the Poisson equation is zero.

2.3 Approximation of the Coulomb potential function

In their paper, Good et al. [34] approximated the r^{-1} term in the Coulomb potential by a sum over three Gaussian functions:

$$\frac{1}{r} = \sum_{i=1}^{3} \lambda_i \mathrm{e}^{-\sigma_i r^2} \tag{9}$$

where the three (λ_i, σ_i) pairs are (0.3001, 0.0499), (0.9716, 0.5026), and (0.1268, 0.0026 Å^{-2}). A visualization of that approximate function clearly shows that the fit of r^{-1} is acceptable only at distances r that are larger than about 1 Å; the asymptotic behaviour of r^{-1} at $r = 0$ is indeed not satisfied.

A smoothed version can be given by relationships similar to Eqs. 3 and 4:

$$\left(\frac{1}{r}\right)_t = \sum_{i=1}^{3} \frac{\lambda_i}{(1+8\sigma_i t)^{3/2}} \mathrm{e}^{-\frac{\sigma_i}{(1+8\sigma_i t)} r^2} \tag{10}$$

The discrepancies between the Good and Hodgkin's approximation and the original r^{-1} function are strongly reduced when the smoothing factor t differs from zero. This is particularly due to the fact that the infinite asymptotic behaviour at $r = 0$ of function r^{-1} is not present any longer in Eq. 6.

2.4 Atomic property fields

The different 3D atomic property fields $P_i(r)$ selected by Totrov [15], that is, hydrogen bond donor, hydrogen bond acceptor, sp^2 hybridization, lipophilicity, charge, and electronegativity, are represented through Gaussian functions:

$$P_i(\mathbf{r}) = \sum_{a \in A}^{\text{nat}} \varphi_{i,a} e^{-\frac{|\mathbf{r}-\mathbf{R}_a|^2}{\lambda^2}} \tag{11}$$

where $\varphi_{i,a}$ stands for the atomic property i associated with atom a (Table 1 of Ref. [15]), and the so-called effective distance parameter λ is set equal to 1.2 Å [15, 16].

Similar to Eq. 10, a smoothed version of Eq. 11 was implemented as follows:

$$P_{i,t}(\mathbf{r}) = \sum_{a \in A}^{\text{nat}} \frac{\phi_{i,a}}{\left(1 + 8\frac{1}{\lambda}t\right)^{3/2}} e^{-\frac{1/\lambda}{\left(1+8\frac{1}{\lambda}t\right)}|\mathbf{r}-\mathbf{R}_a|^2} \tag{12}$$

When using such a description, most of the hydrogen atoms of the molecular structure are eliminated from the superposition procedure, and only the polar ones, as described by Totrov [15], are kept. In the implementation we set up, a global field descriptor is calculated as a summation over all seven above-mentioned $P_i(r)$ fields.

2.5 Evaluation functions for the alignment of smoothed distribution functions

The selection of a 3D scalar field as a relevant property to determine the similarity degree between two molecules A and B has led to several definitions of similarity measure [2, 38].

The well-known overlap similarity measure is defined by:

$$I_{AB,\text{overlap}} = \int \mathrm{d}\mathbf{r} \rho_{A,t}(\mathbf{r}) \rho_{B,t}(\mathbf{r}) \tag{13}$$

where t is the smoothing degree of the ED.

Another quantity used in our previous work [20] is the so-called Laplacian similarity measure $I_{AB,\text{Laplacian}}$:

$$I_{AB,\text{Laplacian}} = \int \mathrm{d}\mathbf{r} \rho_{A,t}(\mathbf{r}) T \rho_{B,t}(\mathbf{r}) \tag{14}$$

where the operator T is related to the Laplacian operator ∇^2 :

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{15}$$

by $T = -\nabla^2/2$. It has been shown that the $I_{AB,\text{Laplacian}}$ similarity measure can be seen as the overlap integral of the gradient of the ED [2, 39]. In the latter reference, the use of the density gradient in quantum similarity measures is thoroughly described and is evaluated versus the overlap similarity measure.

Similarity measures are involved in several well-known similarity index formulae [2, 38, 40] such as the Carbó (also known as Cosine) index:

$$S_{AB,\text{Carbo}} = \frac{I_{AB}}{\sqrt{I_{AA}}\sqrt{I_{BB}}}, \tag{16}$$

the Hodgkin–Richard (also known as Dice) index:

$$S_{AB,\text{Hodgkin}} = \frac{I_{AB}}{\frac{1}{2}(I_{AA} + I_{BB})}, \tag{17}$$

and the 3D shape Tanimoto similarity index:

Table 1 PDB access codes and net charge (in $|e^-|$) of the molecules considered in the present work

	Structure number												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Elastase	1PPF	4EST											
	0	0											
Endothiapepsin	2ER7	4ER1	4ER2	5ER1	5ER2								
	−1	0	−1	−1	0								
Trypsin	1PPH	1TNH	1TNI	1TNJ	1TNK	1TNL	3PTB						
	+1	+1	+1	+1	+1	+1	+1						
Thermolysin	1THL	1TLP	1TMN	2TMN	3TMN	4TLN	4TMN	5TLN	5TMN	6TMN			
	−2	−2	−2	−1	0	0	−2	−1	−2	−2			
P38	1A9U	1BL6	1BL7	1DI9	1M7Q	1OUK	1OUY	1OVE	1OZ1	1W7H	1W84	1WBO	1YQJ
	0	0	+1	0	+1	+1	+1	+1	0	0	0	0	+1
HRV14	2R04	2R06	2R07	2RM2	2RR1	2RS1	2RS3	2RS5					
	0	0	0	0	0	0	0	0					

$$S_{AB,\text{Tanimoto}} = \frac{I_{AB}}{I_{AA} + I_{BB} - I_{AB}} \quad (18)$$

$S_{AB,\text{Tanimoto}}$ was found to be efficient for the superposition of 3D fields, both in the position space [9] and in the momentum space wherein emphasis is given to the long-range variations of the electron density [41]; it is known to be more sensitive to size differences between two structures. It was also found to be efficient in superpositions of endothiapepsin ligands [20].

2.6 Superposition algorithm

Using a Monte Carlo/Simulated Annealing algorithm (MC/SA), rigid pair alignments were achieved at smoothing degrees t varying between 1.7 and 1.4 bohr^2. The two values were selected after the studies presented in [42, 43], which report topological analyses of PASA and Poisson-based CD distribution functions. Best performances of the approach were observed at values of t where the critical points (local maxima and/or minima) of the smoothed 3D fields correspond to known interaction sites of the ligands [25] or to locations of point charges on amino acids [33, 43].

Our superposition algorithm consists of a sequence of MC loops carried out at linearly decreasing acceptance rates. First of all, the structure to be aligned on the reference molecule is translated to locate the two centres of mass at the same position. At each step of a MC loop, the structure to be aligned is displaced by random translation and rotation steps. The maximal translation and rotation displacements were set equal to 0.5 Å and 0.5 rad, respectively. The new alignment is evaluated using S_{AB} and is accepted only if it is probable enough, that is:

$$p = \mathrm{e}^{-\beta\left(S_{AB}^{\text{old}} - S_{AB}^{\text{new}}\right)} > \xi \quad (19)$$

where ξ is a random number selected between 0 and 1. The parameter β controls the acceptance rate of the MC loop. Twenty values are regularly selected between 0.001 and 0.1. The best alignment, that is, the alignment with the highest S_{AB} value, obtained at a given value of β, is used as the starting point of the MC loop at the next β value. The number of iterations per MC loop was set equal to 10,000.

Starting with the PASA molecular description, several calculations were achieved at $t = 1.7$ and 1.4 bohr^2. It was also considered to work at a given β value and let t vary from a high to a low smoothing value during the MC/SA procedure. This last option did not bring real improvements versus the first option. Indeed, the MC/SA algorithm is built to maximize the similarity degree S_{AB}. Starting with a highly smoothed ED and going to a less well-smoothed ED leads to values of S_{AB} that tend to decrease for a given alignment. Nevertheless, the simultaneous variation of t and β during the MC/SA procedure appeared to provide rather good results, with a high performance on the alignment convergence. The MC/SA parameters to be considered when using the PASA description were finally set to a simultaneous and linear decrease in t and β from 1.7 to 1.4 bohr^2 and from 0.1 to 0.001, respectively. Even if ED contours obtained at $t = 1.7$ and 1.4 bohr^2 are very similar as depicted for endothiapepsin ligand 4 (Fig. 1), letting t vary during the superposition procedure seems to favour the search for a global solution.

When considering the CD and APF descriptors, it appeared that the calculation of the similarity measures should better involve positive integrals (overlap, Laplacian) only, that is, the superposition of negative distributions onto positive distributions (or inversely) should not participate to the total value of S_{AB}. This is done to avoid unfavourable partial alignments of the molecules. Looking at the CD contours illustrated in Fig. 1 for endothiapepsin ligand 4, one clearly distinguishes a larger positive isocontour 0.0002 e$^-$/bohr^3 corresponding to the positive NH_3^+ end (left side of the structures in Fig. 1), that is spread away from the molecular skeleton versus its PASA counterpart. It is also seen, for example, that aromatic groups (right side of the structures in Fig. 1) tend to be surrounded by positive regions while the inner part of the ring itself leads to a negatively charged area.

When using the smoothed Coulomb potential description, one first wished to emphasize the overlap of the electrostatic potential acting in regions of space remote from the molecular skeleton itself, that is, beyond the so-called solvent accessible surface of the molecule. In that case, one first used the overlap similarity measure with the Hodgkin similarity index as Good et al. [44] reported that this last index is sensitive to the magnitude of the 3D descriptor field. It, however, appeared that the Laplacian–Tanimoto combination was more efficient.

To evaluate the success of the alignments, the coordinates of the non-hydrogen atoms of the aligned molecules were compared to the corresponding coordinates of their expected (crystallographic) position. An rmsd value was calculated to quantify this degree of success.

3 Applications and results

All 3D coordinates of the molecular systems studied in this paper were retrieved from the Protein Data Base (PDB) [45]. Table 1 reports the PDB access codes of the various systems. Based on the atomic hybridization states, H atoms were added to the structures with the program VEGA ZZ [46, 47]. Protonation states were considered as found in the literature. All end NH_2 and COOH groups in peptides were systematically ionized. Charges were added with the same

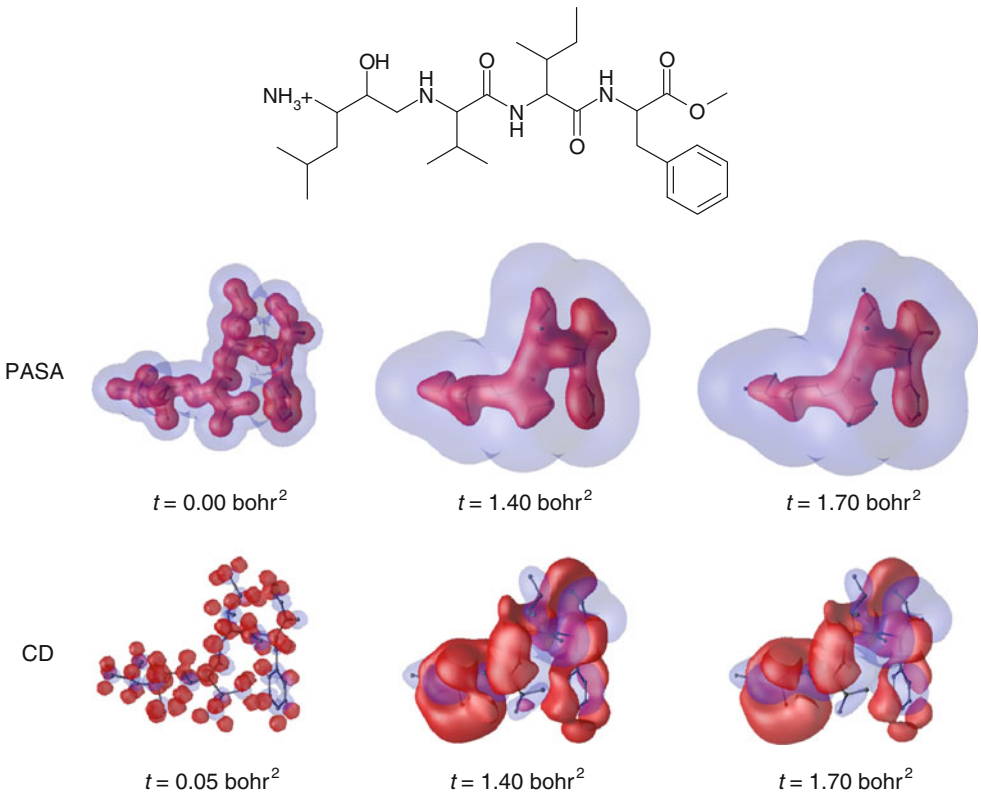

Fig. 1 2D representation and isocontours of the PASA ED (0.0002 in *light blue* and 0.075 $e^-/bohr^3$ in *dark red*) and of the CD (−0.0002 in *light blue* and 0.0002 $e^-/bohr^3$ in *dark red*) of endothiapepsin ligand 4, calculated at various smoothing degrees *t*. The molecular skeleton is displayed using *sticks* (H are not shown for clarity)

program using the Gasteiger–Marsili [48, 49] scheme. 2D representations of all molecular structures studied in this paper are presented in Online Resources 1–6.

Within each family of ligands, all possible pair alignments were carried out, that is, the largest structure on the smallest and inversely. The best solution observed among the two so-obtained was kept.

3.1 Alignment results for the elastase TOMI/DFKi system

The system is particular in that the turkey ovomucoid inhibitor, TOMI (PDB access code 1PPF), is an elastase inhibitor consisting of 56 residues (814 atoms), that is, characterized by a size drastically larger than the difluoroketone inhibitor, DFKi (PDB access code 4EST), with 70 atoms (Online Resource 1). Due to that particularity, it has been the subject of several studies [26–29] regarding their alignment using molecular similarity-based techniques.

The desired alignment of TOMI and DFKi, that is, the expected crystallographic solution, is obtained using the PASA descriptor with the conditions applied throughout this paper, that is, *t* varying from 1.70 to 1.40 $bohr^2$, and with the Laplacian Tanimoto similarity measure and index. The corresponding degree of similarity S_{AB} is equal to 0.0697, and the rmsd value of the TOMI structure is 2.26 Å (Fig. 2a). The low value of S_{AB} is due to the size difference in the two structures, leading to similarity measures that differ by an order of magnitude. In the present case, $I_{AA,\text{Laplacian}} = 21.052$, $I_{BB,\text{Laplacian}} = 1.966$, and $I_{AB,\text{Laplacian}} = 1.450$, at $t = 1.4$ $bohr^2$. The use of the Laplacian similarity measure allows to emphasize the importance of the shape of the molecular skeletons. Another similarity measure choice, like "overlap", will force the overlap between the drug DFKi and the helix of TOMI, that is, a high-density region (Fig. 2b).

All other descriptors did not provide the right alignment under the calculation conditions used in the present work. This suggests that the molecular shape is the main information required to align the molecules.

3.2 Alignment results for the endothiapepsin ligands

Endothiapepsin is a single-chain proteinase of 330 amino acids. 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 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 [50]. The hydrogen bonds that position the inhibitor main chain in the active site cleft are largely conserved from one

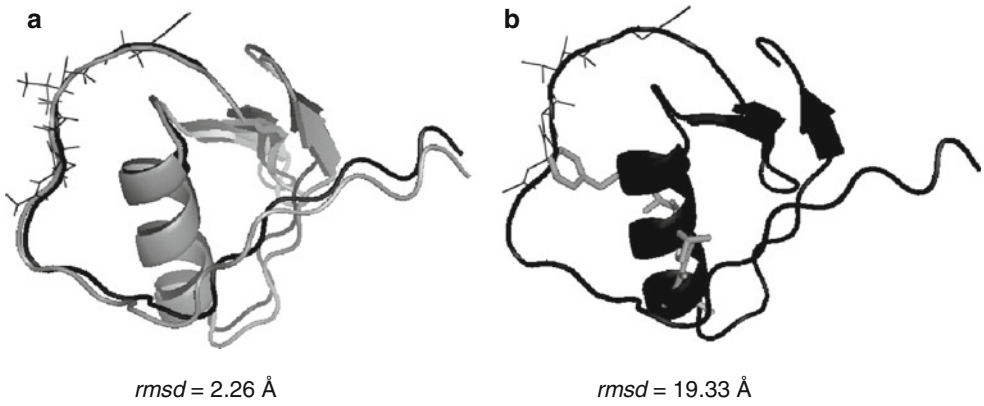

Fig. 2 **a** Superimposition of the crystallographic structure of TOMI (*black ribbon*), DFKi (*black wire*), and the MC/SA orientation of TOMI versus DFKi (*grey*) obtained with the PASA descriptor smoothed at t varying between 1.7 and 1.4 bohr^2 using **a** the Laplacian similarity measure and **b** the overlap similarity measure (H are not shown for clarity)

inhibitor to another, implying that the largest determinants of specificity are the vdW contacts between the enzyme and the ligand side chains. Side chains of the inhibitors can adopt different conformations to compensate for greater or lesser occupation of the neighbouring residues.

The five ligands considered in this work were selected following the work of Lemmen et al. [16] (Online Resource 2). For convenience, the five molecules will be numbered 1–5 further in the text (Table 1).

Alignments achieved using the PASA descriptor together with the Laplacian and Tanimoto similarity evaluators showed that the only problematic case occurred when one superposed ligands 2 and 4 (Table 2). Indeed, the obtained similarity degree S_{AB}, 34 %, is higher than the S_{AB} degree calculated for molecules in the experimental orientation, $S_{AB} = 22$ % at $t = 1.40$ bohr^2, where a partial overlap is observed. With the CD descriptor, this problem is cancelled since $S_{AB\ \text{expected}}$, 25 %, is larger than $S_{AB\ \text{full overlap}}$, 17 %, and a good alignment is obtained between the two ligands 2 and 4 (Table 2). With the CD descriptor, a misalignment remains between ligands 2 and 5. A deeper insight showed that this is due to non-convergence of the algorithm, with $S_{AB\ \text{expected}} = 23$ % at $t = 1.40$ bohr^2, rather than 20 %, as shown in Table 2. Thus, with PASA, only 9 alignments over 10 are successful while one can expect a 100 % success when using the CD descriptor. Lemmen et al. [16] obtained a success rate of 70 %. Besides the use of the PASA and CD distributions functions, the Coulomb potential descriptor did not provide satisfactory alignment results (Table 2). Indeed, only one superposition is characterized by a rmsd value <2 Å, and rmsd values beyond 20 Å suggest that some superposition results completely diverge from the expected ones. The APF descriptor is efficient, except for all alignments that involve ligand 2 (Table 2). This is due to the absence of a negatively charged carboxylate group. Indeed, when present, the two negative oxygen of the carboxylate bear a highly negative charge descriptor value $\varphi_{i,a}$ of -1.5 as well as a highly positive hydrogen bond value, 1.5, in the framework of the APF representation. It is then possible to slightly improve the efficiency of the APF-based alignments, with a success rate of 8 over 10 alignments, by working with the size component of the APF representation only (Table 2).

Thus, one concludes that the orientation of the drug molecules in the binding pocket of the receptor is mainly governed by their shape, a property that is often related to vdW contacts discussed above.

3.3 Alignment results for the trypsin ligands

Trypsins belong to the family of serine proteases and are constituted by about 245 amino acid residues. Their active site contains a serine residue, located at the junction of two β-barrel domains. The seven ligands (Table 1) considered in this work were taken from the work of Chan et al. [18]. Except for the largest ligand, 1PPH, all of them present a rather similar structure (Online Resource 3) that consists in an aromatic group and a positively charged amine group separated by aliphatic linkers of different lengths. All molecules thus bear a net +1 charge. Ligand 1TNI (ligand 3), with a longer linker, assumes a binding mode that is slightly different from the other molecules. The positively substituted phenyl moiety of the larger molecule (ligand 1) is oriented in a similar manner as benzamidine (ligand 7) and faces the negatively charged carboxylate group of residue Asp189 of trypsin while also interacting through hydrogen bonds [51].

None of the alignments involving the largest structure (ligand 1), achieved using either the PASA or the CD descriptor, provided satisfactory results (Online Resources 7 and 8). Indeed, all rmsd values are larger than 5 Å. In such cases, the positively charged N atom of the small ligands is aligned with the SO_2 group of ligand 1 (Fig. 3). This leads to a higher similarity measure due to the larger density distribution of the sulphonate group of ligand 1. With the CD descriptor, two positive regions appear at the level of the $C(NH_2)_2^+$ and SO_2 groups of atoms and

Table 2 Best pair alignment results, in terms of S_{AB} values (%), obtained using the Laplacian Tanimoto MC/SA procedure with various descriptors smoothed at t varying between 1.7 and 1.4 bohr^2 for the five endothiapepsin ligands

	1	2	3	4
PASA				
1	–			
2	35 (0.81)	–		
3	34 (0.16)	40 (0.42)	–	
4	27 (0.72)	**34 (11.14)** $S_{AB\ expected} = 22$	32 (1.42)	–
5	51 (0.47)	49 (0.52)	48 (0.45)	**36 (0.44)**
CD				
1	–			
2	22 (0.73)	–		
3	**31 (1.19)**	29 (0.28)	–	
4	24 (0.93)	**25 (1.65)** $S_{AB\ full\ overlap} = 17$	25 (0.96)	–
5	41 (0.54)	**20 (8.67)** $S_{AB\ expected} = 23$	45 (0.60)	28 (0.80)
Coulomb electrostatic potential				
1	–			
2	8 (10.58)	–		
3	**90 (1.62)**	8 (12.24)	–	
4	1 (41.31)	4 (8.79)	1 (40.00)	–
5	40 (21.00)	10 (8.40)	39 (2.32)	38 (17.64)
APF				
1	–			
2	17 (0.33)	–		
3	60 (0.87)	28 (11.61)	–	
4	**25 (1.24)**	35 (8.92)	31 (1.30)	–
5	60 (0.77)	21 (8.27)	67 (0.61)	27 (0.93)
Size component of the APF				
1	–			
2	36 (0.55)	–		
3	45 (0.37)	44 (0.67)	–	
4	32 (0.74)	30 (9.90)	38 (1.89)	–
5	36 (0.40)	**23 (0.76)**	31 (0.74)	51 (15.99)

The reference molecules are mentioned in the first row; otherwise, results are shown in bold. rmsd values (Å) of the aligned molecules are given in parentheses

misalignments are also observed (Fig. 3). Therefore, the only way to superpose the ligands is to consider a descriptor that involves information well beyond the molecular skeleton, that is, the Coulomb electrostatic potential, or to consider other properties as those in the APF formalism. Figure 3 illustrates that with the Coulomb potential descriptor, there is only one main positive region located around the amidino groups of ligands 1 and 7. The expected alignment can thus be obtained for these two compounds (Fig. 3). Indeed, with that last approach, ones notices that all small molecules tend to be superposed on the correct branch of ligand 1. The Coulomb potential alignments are not ideal, that is, rmsd can be larger than 2 Å, except for the alignment of ligands 1 and 7, with rmsd = 0.36 Å (Online Resource 9). Besides that, the alignment of the small structures versus another did not show significant improvements versus the PASA and CD descriptors. Finally, if one accepts the alignments characterized by rmsd values between 2 and 3 Å, the approach is 100 % successful. In their work, Lemmen et al. [16] also obtained a 100 % success, that is, for each pair of superposed molecules, at least one alignment is correct among the two possible ones. These authors did, however, not consider the largest structure 1PPH in their work. Values

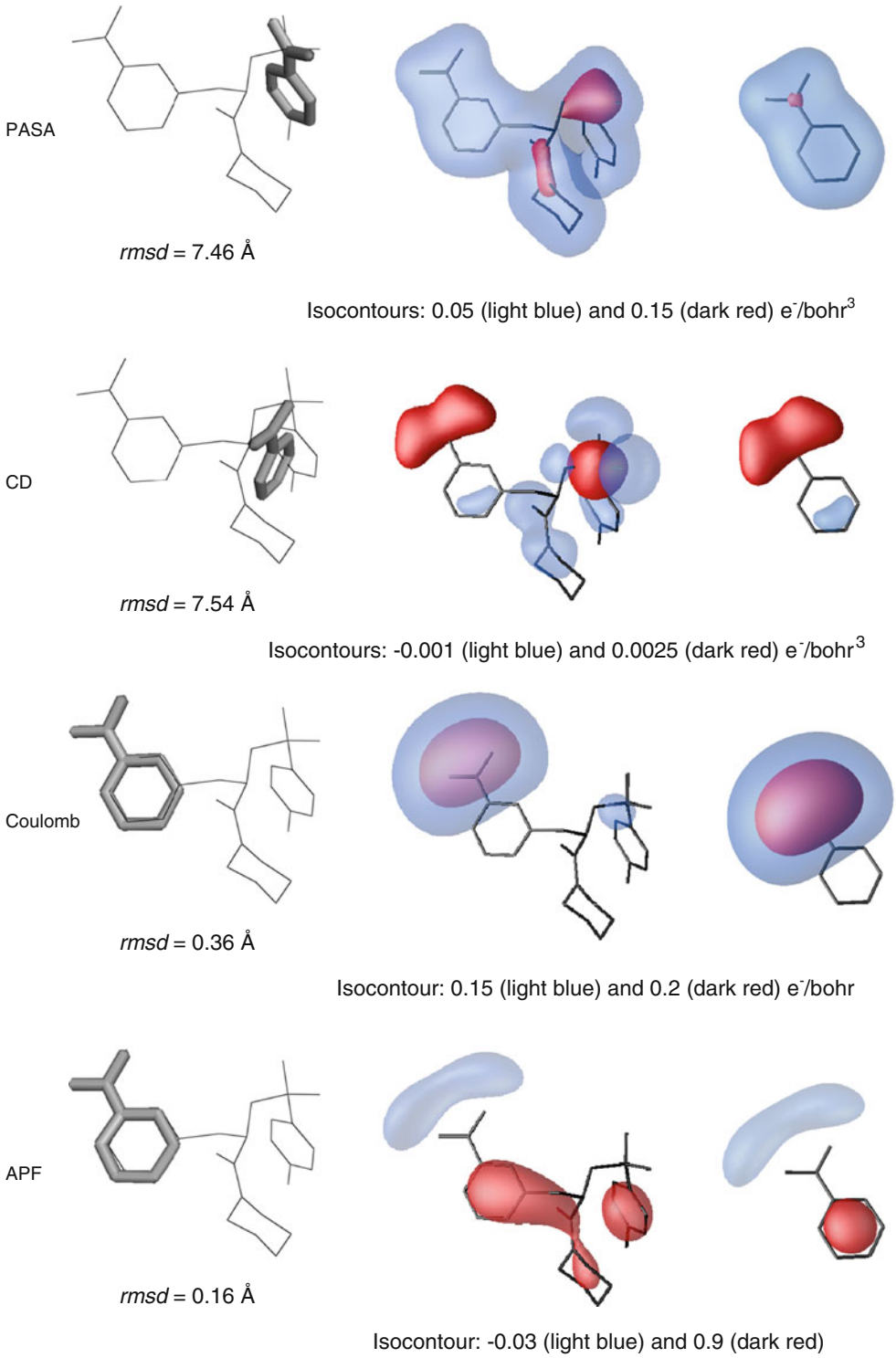

Fig. 3 *Left* Superimpositions of the structures of trypsin ligand 1 (*black wire*) and ligand 7 (*grey stick*) obtained using the MC/SA algorithm with various descriptors smoothed at t varying between 1.7 and 1.4 bohr^2. Isocontours of the PASA ED, CD Coulomb potential, and APF of trypsin ligands 1 (*middle*) and 7 (*right*) calculated at $t = 1.4$ bohr^2. The molecular skeleton is displayed using *sticks* (H are not shown for clarity)

of 80 [18], 73 [32], and 57 % [32] are also reported in the literature. Additional tests carried out using the Coulomb potential descriptor at $t = 1.0$ bohr^2 (Online Resource 9) to determine whether the resolution may lead to lower rmsd values did not show improvements versus the procedure involving a variation of t from 1.70 to 1.40 bohr^2.

The use of the APF descriptor led, as with the Coulomb potential descriptor, to a desired positioning of the aligned ligands (Online Resource 10). The rather high values of rmsd, beyond 2 Å, often characterize shifted molecules, that is, the positive group is well aligned while the aromatic cycle is misaligned as shown for ligands 3 and 7 (Fig. 4). This happens for six alignments, marked "shifted" in Online Resource 10. The alignment of ligands 1 and 6 is characterized by a wrong inversed orientation, with rmsd = 2.21 Å only, and $S_{AB} = 14$ %. A good alignment could be expected at the same similarity degree, that is, $S_{AB} = 14$ %. An additional run was achieved at $t = 0.50$ bohr^2, which tends to show that these shifted positions are not due to the smoothing of the APF field (Online Resource 10). Indeed, nine pairs are misaligned, one inversed orientation is obtained, as well as a completely wrong result for ligands 2 and 7 with rmsd = 5.40 Å.

Superposing the trypsin ligands thus requires a descriptor that is able to differentiate high-density regions of different chemical natures and electric charges by taking into consideration descriptor distributions spread around the molecular skeleton like the Coulomb potential.

3.4 Alignment results for the thermolysin ligands

Thermolysin, a calcium-binding zinc endopeptidase consisting of 316 amino acid residues, involves a pronounced active site cleft formed at the junction of the two lobes characterizing its structure. Ten thermolysin ligands were considered in the present study (Table 1). Two structures, 2TMN and 4TLN, examined by Lemmen et al. [16], were added to the set of eight ligands studied by Chan et al. [18]. Among these ten ligands, 5TMN and 6TMN differ by a small moiety, as shown in Online Resource 4, that is, an O atom replacing a NH group. When bound in the receptor, the molecules are linked to a Zn cation. More precisely, inhibitors 4TLN and 5TLN bind to the receptor with their hydroxamate group complexed to the Zn, while molecules like 1TLP, 2TMN, 4TMN, 5TMN, and 6TMN are coordinated to the Zn by phosphoryl oxygens. Binding to the receptor also occurs through hydrogen bonds with the NH groups of the molecular skeletons. The ionization state of 4TLN and 5TLN was selected following the work of Matthews and coll. [52, 53] who favoured the anionic form of the NHOH moiety. A positive NH_3^+ group is also assumed in structure 4TLN [53]. The structure of 2TMN involves a protonated N atom located next to P, as described by Matthews and coworkers [53, 54]. Finally, O atoms of the phosphoramidate groups in 1TLP, 4TMN, 5TMN, and 6TMN are charged as represented in the work of Gresh et al. [55].

The use of the PASA descriptor (Online Resource 11) provides an overall success rate, 27 over 45 alignments, that is slightly less good than the one obtained with the CD

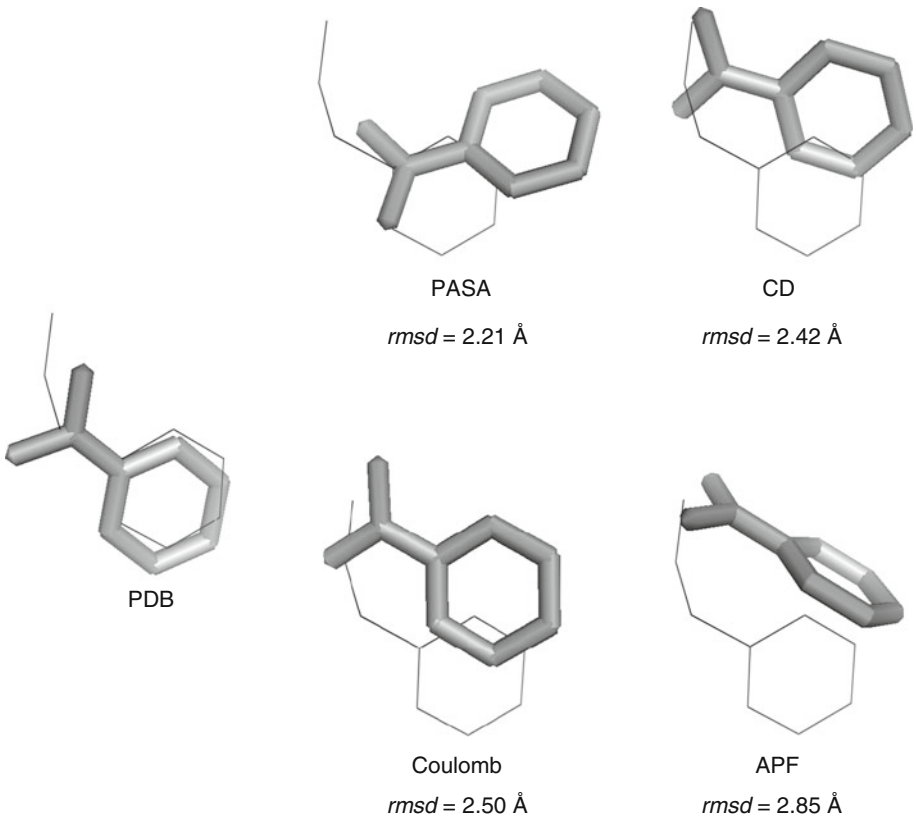

Fig. 4 Superimposition of the structures of trypsin ligand 3 (*black wire*) and ligand 7 (*grey stick*) in its crystallographic orientation (PDB), and in the MC/SA orientations obtained using the PASA ED, CD, Coulomb potential, and APF descriptors smoothed at t varying between 1.7 and 1.4 bohr^2 (H are not shown for clarity)

descriptor, that is, 30 good solutions (Online Resource 12). All rmsd values of the successful alignments are lower than 2 Å, except for the superimposition of ligands 6 and 10 with the PASA descriptor, rmsd = 2.62 Å, and the superposition of ligands 7 and 10 with the CD descriptor, rmsd = 2.66 Å. Proper alignments are expected with unchanged S_{AB} values between ligands 2 and 6 using the PASA and CD descriptors. The main difference between the PASA and CD-based results lies in the type of alignments that were achieved with success. For example, all nine alignments involving ligand 8 led to expected results with the CD descriptor, while only three were obtained as desired with PASA. Combining both sets of results leads to a total success rate of 36 over 45 solutions, that is, 80 %. This may suggest that a combination of shape- and charge-dependent descriptors needs to be envisaged for further work with those ligands. The case of ligands 4 and 10 is depicted in Fig. 5, which illustrates that the wrong alignment observed with the CD descriptor occurs due to the large negative region of ligand 4 located at the level of PO_3^{2-}. An additional run carried out with a NH group, replacing the NH_2^+ moiety, did not modify at all the alignment results. The use of the smoothed Coulomb potential (Online Resource 13) did not bring any overall improvement over PASA, especially for all superpositions involving ligand 4, the smallest one, and ligand 5; 17 good solutions, characterized by rmsd < 2 Å, were obtained. Seven alignments correspond to shifted molecules versus their expected orientation. As for the PASA descriptor, the APF one (Online Resource 14) leads to mixed results, with a success rate of 20 over 45 alignments. These good solutions are all characterized by rmsd < 2.5 Å. For ligands 6 and 8, we have to mention that an additional APF type was considered for the negatively charged N atoms of the hydroxamate groups occurring in both ligands. The selected parameters were −0.5 for hydrogen bond donor, 1.5 for hydrogen bond acceptor, 0.0 for sp^2 hybridization, −1.0 for lipophilicity, 0.0 for size, −1.5 for charge, and −1.5 for electronegativity.

Only molecules 1–3, 5, and 7–10 were considered in the work by Chan et al. [18], while molecules 2–9 were studied by Lemmen et al. [16]. If one restricts our analyses to the eight structures involved in each of those previous studies, one gets a maximal success rates of 82 % (23 alignments over 28) and 64 % (18 alignments over 28) with the CD descriptor that are of the same order of magnitude as the literature values of 93 % [18] and 61 % [16], respectively. Only alignments with rmsd < 2 Å were considered, as in the literature [16, 18]. On the whole, most of the wrong superposition results obtained with the CD descriptor (Online Resource 12) involve molecules 4–6, that is, the smallest structures. As already discussed, structure 8 that is a bit larger in size is always well aligned with CD. The particular case of ligands 1 and 8 is illustrated in Fig. 6. The use of the CD descriptor allowed to correctly superpose the functional moieties that are expected to bind to the Zn ion, that is, the carboxylate and the hydroxamate groups, respectively, and the remaining C=O functions (Fig. 6b). On the contrary, PASA tends to superpose the –NH–CH_2–CH_2–OH tail of molecule 8 with the fused rings of ligand 1 (Fig. 6a) to maximize S_{AB}.

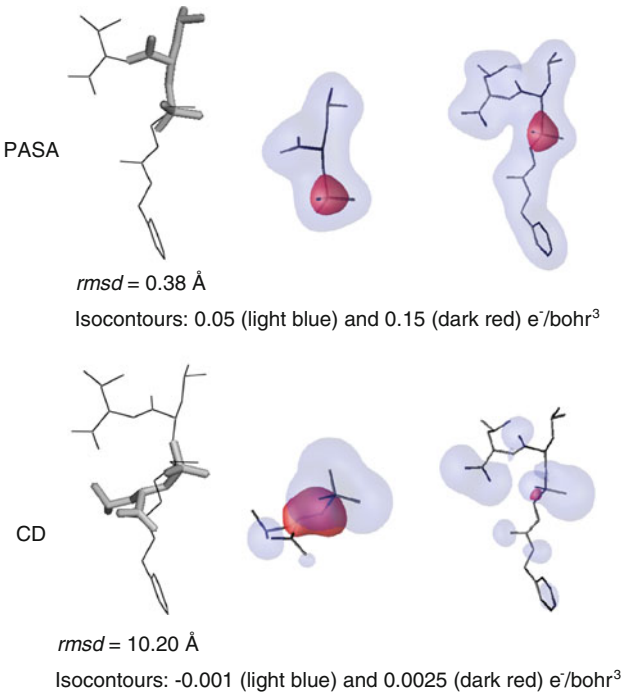

Fig. 5 *Left* Superimpositions of the structures of thermolysin ligand 4 (*grey stick*) and ligand 10 (*black wire*) obtained using the MC/SA algorithm with various descriptors smoothed at t varying between 1.7 and 1.4 $bohr^2$. Isocontours of the PASA ED and CD of thermolysin ligands 4 (*middle*) and 10 (*right*) calculated at $t = 1.4$ $bohr^2$. The molecular skeleton is displayed using *sticks* (H are not shown for clarity)

A careful analysis of the molecular structures is thus required to select a descriptor for this family of molecules in order to determine whether an emphasis is to be given on isosterism or on the electric charge of the functional groups.

3.5 Alignment results for the p38 ligands

Among the four members of the p38 MAP kinase family, the most studied isoform is p38α, a target for anti-inflammatory drugs. The primary sequence reported in the PDB consists of 360–379 amino acid residues forming secondary structures of types α and β. Thirteen ligands (Table 1), able to bind in the ATP-binding pocket located between the N- and C-terminal lobes of p38α, were studied following the work of Chan et al. [18]. The 2D structures are reported

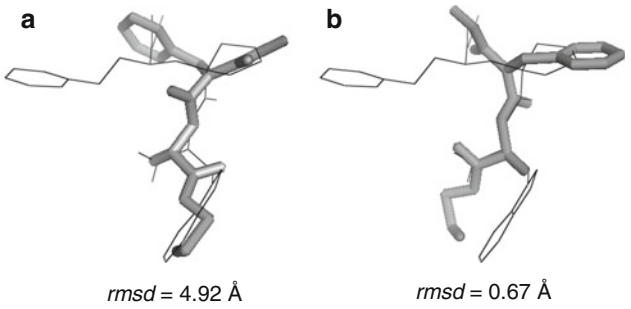

Fig. 6 Superimposition of the structures of thermolysin ligand 1 (*black wire*) and ligand 8 (*grey stick*) obtained using the MC/SA algorithm with the **a** PASA and **b** CD descriptors smoothed at t varying between 1.7 and 1.4 bohr^2 (H are not shown for clarity)

in Online Resource 5; the protonation states were selected according to the work of Chan et al. [18]. The p38 ligands interact with the receptor mainly through hydrogen bonding, π-stacking, and hydrophobic interactions. A study of the crystalline complexes shows that the ligands bind in different arrangements in the same binding pocket, as illustrated in Fig. 6 of [18]. Ligand 1WBO is particularly small as it results from fragment-based lead discovery studies, and all four ligands 5–8, that is, 1M7Q, 1OUK, 1OUY, and 1OVE bind in a similar way. Additionally, all those four ligands contain halogen atoms, their pyridyl/pyrimidyl protonated nitrogen form a hydrogen bond with the main chain NH of Met109, and their aryl substituent occupies a hydrophobic pocket [57]. Ligand 1DI9 binds in a mode different from the others, but still interacts with Met109 and Thr106 [58]. The seven remaining ligands, that is, ligands 1–3, 9–11, and 13, bind in a similar fashion; the inhibitors interact with the receptor through hydrogen bonds between their pyridyl/pyrimidyl moieties and Met109, and their respective aromatic groups occupy a lipophilic pocket involving Thr106 [59–62]. Only some of them, ligands 1–3 and 9, are halogenated. A hydrogen bond may be involved with Lys58 [63]. A superposition of the thirteen ligands in their crystallographic orientation is given in Fig. 7 to classify them according to their binding fashion. Such different binding patterns made the superposition between members of the different binding families rather unsuccessful. For example, success rate values of 43 [18] and 27 % [32] were reported in the literature.

Similarity indices and rmsd values are given in Online Resources 15–18 for all alignments obtained with the PASA, CD, Coulomb potential, and APF descriptors, respectively. Alignments between molecules that bind in a similar fashion are highlighted in light and dark grey in the Tables of the Online Resources and the corresponding total number of expected good alignments is 27. With the PASA descriptor, an overall success rate of 37 % (29 alignments over 78) was obtained. When considering molecules in similar binding families, seventeen alignments over 27 were satisfactory, that is, 63 %, all with rmsd below 2.1 Å. With the CD and Coulomb potential descriptors, only eight good alignments were obtained. Moreover, those eight alignments are characterized by a larger rmsd limit value, that is, 2.9 Å. One can additionally observe that with PASA, three alignments did not converge towards the expected solutions. Indeed, proper alignments can be obtained for ligands 2 and 7 ($S_{AB} = 41$ %), 6 and 8 ($S_{AB} = 51$ %), and 7 and 11 ($S_{AB} = 36$ %). This brings the success rate of the PASA-based superposition procedure to a value of 67 % for the alignments obtained using molecules of the same binding families. In comparison, the corresponding success rate reached by Chan et al. [18] is 78 %. In both Chan et al.'s work and in ours, a 100 % success is reached for the family of structures 5–8. With the APF descriptor, only eight expected alignments are obtained, all with rmsd < 2 Å. Beyond that value, the rmsd value cannot be associated with a good alignment. For instance, the incorrect superposition of ligands 6 and 13, illustrated in Fig. 8, is characterized by rmsd = 2.77 Å.

According to the superposition results, the molecular similarity in the ligands of the p38 MAP kinases is mainly shape-dependent. Most of these ligands involve halogen atoms, which guide the alignment procedure. For ligands

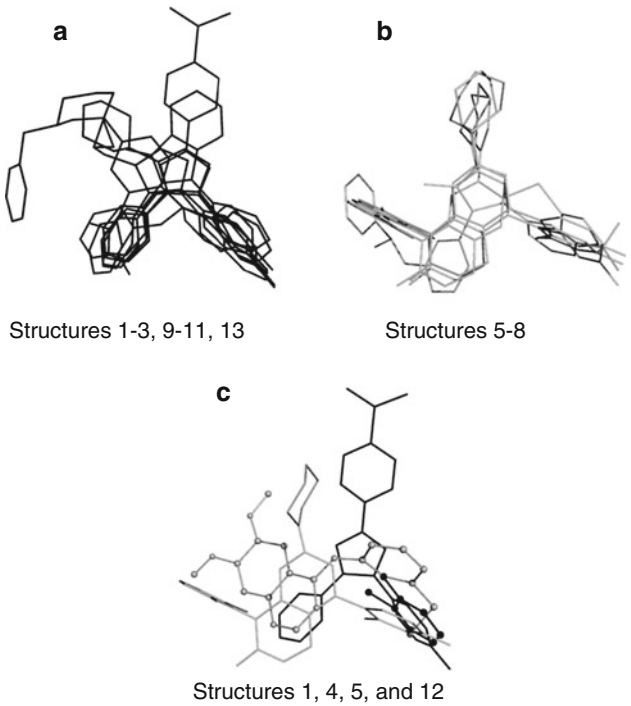

Fig. 7 Superimpositions of the p38 ligands according to their binding mode: **a** ligands *1–3*, *9–11*, and *13*; **b** ligands *5–8*; **c** ligands *5* (*grey*) and *12* (*black ball*-and-*stick*) are displayed together with ligands 1 (*black*) and *4* (*grey ball*-and-*stick*) for comparison with the two binding families (**a**) and (**b**) (H are not shown for clarity)

5–8, the CD descriptor also provides acceptable results, due to the presence of NH_2^+ moieties in all four molecules, and hence can be adequately superposed based on that charged group.

3.6 Alignment results for the HRV14 ligands

Eight antiviral compounds binding to the coating protein of human rhinovirus 14 (Table 1) were considered as in the work of Lemmen et al. [16]. The RHV14 coat protein has four subunits named VP1–VP4 with 289, 262, 236, and 68 amino acid residues, respectively. The HRV14 viral capsid consists of 60 copies of each subunit. VP1 to VP3 are located at the viral capsid surface, while VP4 is buried deeper in the virion, close to the capsid/RNA interface. All ligands are rather extended molecules composed of heterocycles at both ends, which are separated by an aliphatic chain and an aromatic group that act as linkers (Online Resource 6). The HRV14 inhibitors show two distinct binding modes [56] that differ in the orientation of the ligand. A reverse binding mode is observed for ligands 2RM2, 2RR1, 2RS1, and 2RS3 that are characterized by a seven carbon long linker and two methylated five-membered rings, versus 2R04, 2R06, 2R07, and 2RS5. Those last four molecules are all characterized by a shorter linker, constituted of five C atoms, with no or only one methylated five-membered ring. The binding pocket of the receptor is mainly hydrophobic as shown in Fig. 2 of [64] and is essentially composed of residues of VP1 that form a β-barrel.

As observed in the studies of Lemmen et al. [16] and Tervo et al. [56], the use of molecular fields does not always allow to detect reverse orientations. Nevertheless, among all descriptors used in the present work, PASA is able to adequately superimpose ligands of the two orientations. The particular cases of ligands 2–3 and 2–4 are presented in Fig. 9 for the various descriptors. When using the PASA descriptor (Online Resource 19), all alignments between molecules binding in a similar orientation, that is, molecules 1–3 and 8, and molecules 4–7, are successful and are all characterized by rmsd < 1 Å. It is noteworthy to mention that the alignment of molecules binding in a different orientation is also satisfying, as illustrated in Fig. 9 for ligands 2 and 4, except for four cases: ligands 1 and 4, 3 and 4, 3 and 7, and 4 and 8, the success rate being 12 over 16. With the charge-dependent descriptors, like CD (Online Resource 20), Coulomb potential (Online Resource 21), and APF (Online Resource 22), almost all alignments of molecules binding in the same orientation are correct, with rmsd < 1 Å, while no correct alignments between ligands of two inverse orientations are found, that is, rmsd > 10 Å, as also observed by Tervo et al. [56]. An illustration that emphasizes the similarity of the electrostatic potential isocontours explaining the incorrect alignments of ligands of different binding orientations can be found in Fig. 2 of Ref. [56]. The superposition problem occurring between ligands 2 and 4 is not strictly dependent on the shape of the molecules. Indeed, the use of the size component of the APF descriptor only did not provide any good results either. Thus, the nature of the atoms constituting the molecules should also be considered.

Regarding the performance of other approaches, Lemmen et al. [16] obtained a success rate of 100 % when aligning molecules of the same orientations only. Overall values of about 50 % were obtained by Chan et al. [18] and the programs ROCS and FLEXS [32].

One thus concludes that the PASA descriptor, which is able to involve the shape and size of the molecules, as well as the chemical nature of the atoms, is essential for aligning as desired the molecules of the rhinovirus ligand family, and charge effect should be avoided.

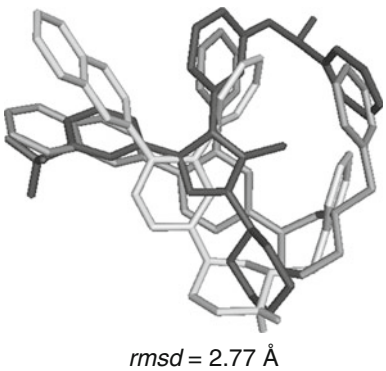

Fig. 8 Superimposition of the structures of p38 ligand 6 (*black stick*) and ligand 13 in its crystallographic orientation (*dark grey stick*), and in the MC/SA orientation obtained using the APF descriptor smoothed at t values between 1.7 and 1.4 bohr^2 (*light grey stick*) (H are not shown for clarity)

4 Conclusions

In the present work, a Monte Carlo/Simulated Annealing rigid superposition algorithm was applied to six families of drug molecules, that is, elastase inhibitors, and ligands of endothiapepsins, trypsins, thermolysins, p38 MAP kinases, and rhinovirus, for which various alignment problems were reported in the literature.

All molecules were described using each of the four following smoothed molecular fields, that is, the promolecular atomic shell approximation (PASA) of the full electron density (ED) [35], a charge density (CD) calculated using the Poisson equation [33], the Coulomb electrostatic potential [34], and the Atomic Property Fields (APF) described by Totrov [15].

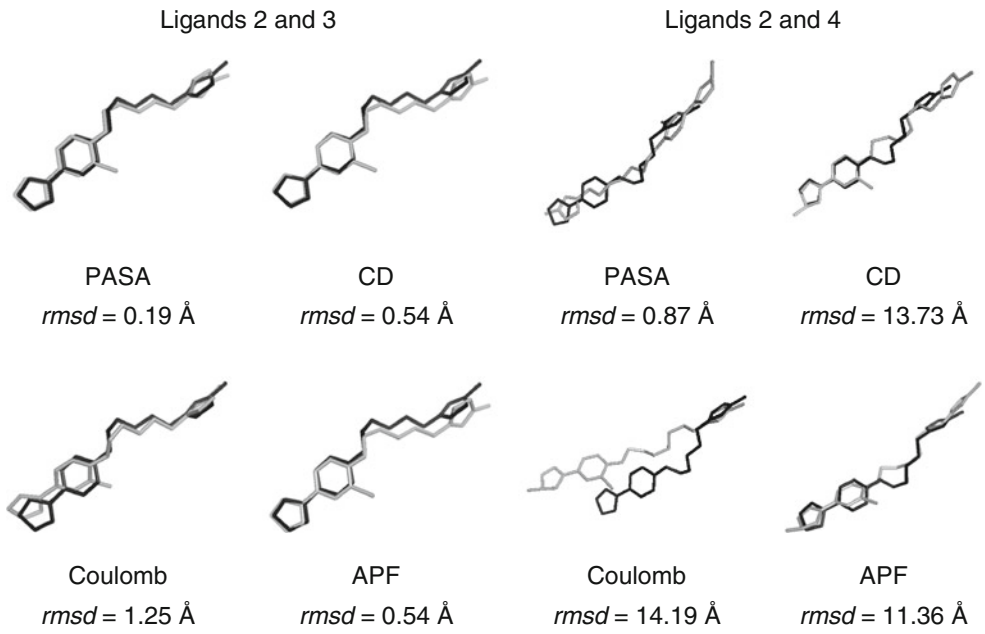

Fig. 9 Superimposition of the structures of HRV14 ligand 2 (*black stick*) and (*left*) ligand 3 (*grey stick*), and (*right*) ligand 4 (*grey stick*) in their MC/SA orientation obtained using the PASA ED, CD, Coulomb potential, and APF descriptors smoothed at t varying between 1.7 and 1.4 bohr^2 (H are not shown for clarity)

All descriptor fields were smoothed to lower the number of local solutions. An additional consequence of that smoothing resides in the levelling of the similarity degree values. This involves that in some cases, several different alignment solutions may occur with the same similarity degrees as for the desired alignments. The smoothing degree was selected following previous studies wherein it was shown that the topological properties of the promolecular ED and the CD distribution functions could be related to molecular features such as pharmacophore elements and/or amino acid residues in larger biomolecules [20, 33]. It was observed that a simultaneous change in the "temperature" during the simulated annealing process, carried out together with a decrease in the smoothing degree, favoured the convergence to global solutions.

With this work, it was first noticed that the use of a rmsd value to evaluate the pair-wise alignments is appropriate for selected ranges of values. All alignments characterized by rmsd smaller or equal to 2 Å correspond to expected solutions. Beyond 3 Å, all can be considered as inadequate. Between 2 and 3 Å, they mostly correspond to displaced alignments but, in some less frequent cases, to false positives. For large peptide structures, such as the elastase ligand TOMI, it can correspond to a good solution.

On the whole, the PASA descriptor field appeared to be the best choice to superpose amino acid sequences, like the two elastase inhibitors, and molecules interacting mainly through van der Waals contacts with their receptor. This was especially clear for molecules of the elastase, endothiapepsin, and rhinovirus ligand families. Alignments of the ligands of the p38 MAP kinase family provided slightly less successful results versus the other methods proposed in the literature. When the shape of the molecules is not the essential component to consider in the description of the molecules, the use of property fields such as the CD and the Coulomb potential can bring a real improvement versus PASA, especially in the families of trypsin and thermolysin inhibitors. A 100 % success was obtained for trypsin when using the smoothed Coulomb potential as a descriptor.

It is thus considered that the descriptor to select for the alignments is strongly dependent upon the nature of the interactions between the drug molecules and their receptor. Additional parameters that are difficult to control during superpositions, but that may affect the results if considered, are flexibility, hydration state, presence of metallic ions or clusters, etc. In addition, working with ligands of very different size, or which partly overlap in the receptor, is always a challenge, but it appeared that these difficulties may also be overcome by an adequate choice of the descriptor, as shown, for example, with the elastase and endothiapepsin ligands. Working with several descriptors may also allow to cumulate adequate alignment information.

To extend and/or improve the superposition results, different strategies could be considered. Besides the inclusion of flexibility, one approach consists in aligning at least three molecules at a time, as discussed by Mestres et al. [65]. Such considerations would, however, lead to additional local solutions to the superposition problems.

One also might to combine several similarity degrees, obtained using different descriptors such as PASA, CD, Coulomb potential, as done in the present work by using the seven APF descriptors of Totrov [15], and by Mestres et al. [65].

5 Online resources

The 2D structure and protonation states of all molecules considered in the present work are given in the Online Resources, as well as the alignment results, in terms of similarity degrees S_{AB} and rmsd values, for the trypsins, thermolysins, p38 MAP kinases, and rhinovirus ligands.

Acknowledgments The authors thank the reviewers for their comments. They also acknowledge L. Piela for fruitful discussions, the support of the F.R.S.–F.R.F.C. (convention no. 2.4.617.07.F), and the ''Facultés Universitaires Notre-Dame de la Paix'' (FUNDP) for the use of the Interuniversity Scientific Computing Facility (ISCF) Center.

References

1. Maggiora GM, Shanmugasundaram V (2004) Methods in molecular biology. In: Bajorath J (ed) Chemoinformatics: concepts, methods, and tools for drug discovery, vol 275. Humana Press, Totowa
2. Bultinck P, Gironés X, Carbó-Dorca R (2005) In: Lipkowitz KB, Larter R, Cundari TR (eds) Reviews in computational chemistry, vol 21. Wiley-VCH, Hoboken
3. Carbó-Dorca R, Besalú E, Mercado LD (2011) Communications on quantum similarity, Part 3: a geometric-quantum similarity molecular superposition algorithm. J Comput Chem 32:582–599
4. Connolly ML (1996) NetSci's science center: computational chemistry 1996. http://www.netsci.org/Science/Compchem/feature14.html. Accessed 11 Jan 2012
5. Good AC, Richards WG (1993) Rapid evaluation of shape similarity using Gaussian functions. J Chem Inf Comput Sci 33:112–116
6. Grant JA, Pickup BT (1995) A Gaussian description of molecular shape. J Phys Chem 99:3503–3510
7. Grant JA, Gallardo MA, Pickup BT (1996) A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. J Comput Chem 17:1653–1666
8. Grant JA, Pickup BT (1997) Gaussian shape methods. Comput Simul Biomol Syst 3:150–176
9. Nicholls A, MacCuish NE, MacCuish JD (2004) Variable selection and model validation of 2D and 3D molecular descriptors. J Comput Aided Mol Des 18:451–474
10. Haigh JA, Pickup BT, Grant JA, Nicholls A (2005) Small molecules shape-fingerprints. J Chem Inf Model 45:673–684
11. Maggiora GM, Rohrer DC, Mestres J (2001) Comparing protein structures: a Gaussian-based approach to the three-dimensional structural similarity of proteins. J Mol Graph Model 19:168–178
12. Mestres J, Rohrer DC, Maggiora GM (1997) MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches. J Comput Chem 18:934–954
13. Duncan BS, Olson A (1993) Shape analysis of molecular surfaces. Biopolymers 33:231–238
14. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
15. Totrov M (2008) Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. Chem Biol Drug Des 71:15–27
16. Lemmen C, Lengauer T, Klebe G (1998) FLEXS: a method for fast flexible ligand superposition. J Med Chem 41:4502–4520
17. Proschak E, Rupp M, Derksen S, Scneider G (2008) Shapelets: possibilities and limitations of shape-based virtual screening. J Comput Chem 29:108–114
18. Chan SL, Labute P (2010) Training a scoring function for the alignment of small molecules. J Chem Inf Model 50:1724–1735
19. Molecular Operating Environment (MOE), 2011.10 (2011) Chemical Computing Group Inc., Montreal
20. Leherte L (2006) Similarity measures based on Gaussian-type promolecular electron density models: alignment of small rigid molecules. J Comput Chem 27:1800–1816
21. Kostrowicki J, Piela L, Cherayil BJ, Scheraga HA (1991) Performance of the diffusion equation method in searches for optimum structures of clusters of Lennard-Jones atoms. J Phys Chem 95:4113–4119
22. Glick M, Robinson DD, Grant GH, Richards WG (2002) Identification of ligand binding sites on proteins using a multi-scale approach. J Amer Chem Soc 124:2337–2344
23. Glick M, Grant GH, Richards WG (2002) Docking of flexible molecules using multiscale ligand representations. J Med Chem 45:4639–4646
24. Leherte L (2001) Application of multiresolution analyses to electron density maps of small molecules: critical point representations for molecular superposition. J Math Chem 29:47–83
25. Leherte L, Meurice N, Vercauteren DP (2005) Influence of conformation on the representation of small flexible molecules at low resolution: alignment of endothiapepsin ligands. J Comput Aided Mol Des 19:525–549
26. Masek BB, Merchant A, Matthew JB (1993) Molecular skins: a new concept for quantitative shape matching of a protein with its small molecule mimics. Proteins 17:193–202
27. Perkins TDJ, Mills JEJ, Dean PM (1995) Molecular surface-volume and property matching to superpose flexible dissimilar molecules. J Comput Aided Mol Des 9:479–490
28. Poirrette AR, Artymiuk PJ, Rice DW, Willett P (1997) Comparison of protein surfaces using a genetic algorithm. J Comput Aided Mol Des 11:557–569
29. Robinson DD, Lyne PD, Richards WG (2000) Partial molecular alignment via local structure analysis. J Chem Inf Comput Sci 40:503–512
30. Klebe G, Mietzner T, Weber F (1994) Different approaches toward an automatic structural alignment of drug molecules: applications to sterol mimics, thrombin and thermolysin inhibitors. J Comput Aided Mol Des 8:751–778
31. Vieth M, Hirst JD, Brooks CL III (1998) Do active site conformations of small ligands correspond to low free-energy solution structures? J Comput Aided Mol Des 12:563–572
32. Chen Q, Higgs RE, Vieth M (2006) Geometric accuracy of three-dimensional molecular overlays. J Chem Inf Model 46: 1996–2002
33. Leherte L, Vercauteren DP (2011) Charge density distributions derived from smoothed electrostatic potential functions: design of protein reduced point charge models. J Comput Aided Mol Des 25:913–930
34. Good AC, Hodgkin EE, Richards WG (1992) Utilization of Gaussian functions for the rapid evaluation of molecular similarity. J Chem Inf Comput Sci 32:188–191

35. Amat L, Carbó-Dorca R (2000) Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. J Chem Inf Comput Sci 40:1188–1198
36. Amat L, Carbó-Dorca R (1997) Quantum similarity measures under atomic shell approximation: first order density fitting using elementary Jacobi rotations. J Comput Chem 18:2023–2039. http://iqc.udg.es/cat/similarity/ASA/funcset.html. Accessed 12 Jan 2012
37. Hart RK, Pappu RV, Ponder JW (2000) Exploring the similarities between potential smoothing and simulated annealing. J Comput Chem 21:531–552
38. Robert D, Carbó-Dorca R (1998) A formal comparison between molecular quantum similarity measures and indices. J Chem Inf Comput Sci 38:469–475
39. Carbó-Dorca R, Mercado LD (2010) Commentaries on quantum similarity (1): density gradient quantum similarity. J Comput Chem 31:2195–2212
40. Maggiora GM, Petke JD, Mestres J (2002) A general analysis of field-based molecular similarity indices. J Math Chem 31:251–270
41. Cooper DL, Allan NL (1995) In: Carbó R (ed) Molecular similarity and reactivity: from quantum chemical to phenomenological approaches. Kluwer, Dordrecht
42. Leherte L (2004) Hierarchical analysis of promolecular full electron-density distributions: description of protein structure fragments. Acta Crystallogr Sect D 60:1254–1265
43. Leherte L, Vercauteren DP (2009) Coarse point charge models for proteins from smoothed molecular electrostatic potentials. J Chem Theory Comput 5:3279–3298
44. Good AC, Peterson SJ, Richards WG (1993) QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. J Med Chem 36(1993):2929–2937
45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242. http://www.rcsb.org/pdb. Accessed 12 Jan 2012
46. Pedretti A, Villa L, Vistoli G (2002) VEGA: a versatile program to convert, handle, and visualize molecular structure on Windows-based PCs. J Mol Graph 21:47–49
47. Pedretti A, Villa L, Vistoli G (2004) VEGA—an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. J Comput Aided Mol Des 18:16–173. http://www.vegazz.net/. Accessed 12 Jan 2012
48. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. Tetrahedron 36:3219–3222
49. Marsili M, Gasteiger J (1981) π Charge distribution from molecular topology and π orbital electronegativity. Croat Chem Acta 53:601–614
50. Bailey D, Cooper JB (1994) A structural comparison of 21 inhibitor complexes of the aspartic proteinase from *Endothia parasitica*. Protein Sci 3:2129–2143
51. Turk D, Stürzebecher J, Bode W (1991) Geometry of binding of the Nα-tosylated piperidines of *m*-amidino-, *p*-amidino- and *p*-guanidino phenylalanine to thrombin and trypsin—X-ray crystal structures of their trypsin complexes and modeling of their thrombin complexes. FEBS 287:133–138
52. Holmes MA, Matthews BW (1981) Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis. Biochemistry 20:6912–6920
53. Matthews BW (1988) Structural basis of the action of thermolysin and related zinc peptidases. Acc Chem Res 21:333–340
54. Tronrud DE, Monzingo AF, Matthews BW (1986) Crystallographic structural analysis of phosphoramidates as inhibitors and transition-state analogs of thermolysin. Eur J Biochem 157:261–268
55. Gresh N, Roques BP (1997) Thermolysin-inhibitor binding: effect of the $His^{231} \rightarrow$ Ala mutation on the relative affinities of thiolate versus phosphoramidate inhibitors—A model theoretical investigation incorporating a continuum reaction field hydration model. Biopolymers 41:145–164
56. Tervo AJ, Rönkkö T, Nyrönen TH, Poso A (2005) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. J Med Chem 48:4076–4086
57. Stelmach JE, Liu L, Patel SB, Pivnichny JV, Scapin G, Singh S, Hop CECA, Wang Z, Strauss JR, Cameron PM, Nichols EA, O'Keefe SJ, O'Neill EA, Schmatz DM, Schwartz CD, Thompson CM, Zaller DM, Doherty JB (2003) Design and synthesis of potent, orally bioavailable dihydroquinazolinone inhibitors of p38 MAP kinase. Bioorg Med Chem Lett 13:277–280
58. Shewchuk L, Hassell A, Wisely B, Rocque W, Holmes W, Veal J, Kuyper LF (2000) Binding mode of the 4-anilinoquinazoline class of protein kinase inhibitor: X-ray crystallographic studies of 4-anilinoquinazolines bound to cyclin-dependent kinase 2 and p38 kinase. J Med Chem 43:133–138
59. Wang Z, Canagarajah BJ, Boehm JC, Kassicà S, Cobb MH, Young PR, Abdel-Meguid S, Adams JL, Goldsmith EJ (1998) Structural basis of inhibitor selectivity in MAP kinases. Structure 6:1117–1128
60. Gill AL, Frederickson M, Cleasby A, Woodhead SJ, Carr MG, Woodhead AJ, Walker MT, Congreve MS, Devine LA, Tisi D, O'Reilly M, Seavers LCA, Davis DJ, Curry J, Anthony R, Padova A, Murray CW, Carr RAE, Jhoti H (2005) Identification of novel p38α MAP kinase inhibitors using fragment-based lead generation. J Med Chem 48:414–426
61. Tamayo N, Liao L, Goldberg M, Powers D, Tudor YY, Yu V, Wong LM, Henkle B, Middleton S, Syed R, Harvey T, Jang G, Hungate R, Dominguez C (2005) Design and synthesis of potent pyridazine inhibitors of p38 MAP kinase. Bioorg Med Chem Lett 15:2409–2413
62. Perry JJP, Harris RM, Moiani D, Olson AJ, Tainer JA (2009) p38α MAP kinase C-terminal domain binding pocket characterized by crystallographic and computational analyses. J Mol Biol 391:1–11
63. Trejo A, Arzeno H, Browner M, Chanda S, Cheng S, Comer DD, Dalrymple SA, Dunten P, Lafargue J, Lovejoy B, Freire-Moar J, Lim J, Mcintosh J, Miller J, Papp E, Reuter D, Roberts R, Sanpablo F, Saunders J, Song K, Villasenor A, Warren SD, Welch M, Weller P, Whiteley PE, Zeng L, Goldstein DM (2003) Design and synthesis of 4-azaindoles as inhibitors of p38 MAP kinase. J Med Chem 46:4702–4713
64. Hadfield AT, Oliveira MA, Kim KH, Minor I, Kremer MJ, Heinz BA, Shepard D, Pevear DC, Rueckert RR, Rossmann MG (1995) Structural studies on human rhinovirus 14 drug-resistant compensation mutants. J Mol Biol 253:61–73
65. Mestres J, Rohrer DC, Maggiora GM (1999) A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. J Comput Aided Mol Des 13:79–93