J Comput Aided Mol Des (2006) 20:227–236
DOI 10.1007/s10822-006-9052-4

ORIGINAL PAPER

BRUTUS: Optimization of a grid-based similarity function for rigid-body molecular superposition. II. Description and characterization

Toni Rönkkö · Anu J. Tervo · Jussi Parkkinen · Antti Poso

Received: 12 April 2006 / Accepted: 14 June 2006 / Published online: 20 July 2006

Abstract Finding novel lead molecules is one of the primary goals in early phases of drug discovery projects. However, structurally dissimilar compounds may exhibit similar biological activity, and finding new and structurally diverse lead compounds is difficult for computer algorithms. Molecular energy fields are appropriate for finding structurally novel molecules, but they are demanding to calculate and this limits their usefulness in virtual screening of large chemical databases. In our approach, energy fields are computed only once per superposition and a simple interpolation scheme is devised to allow coarse energy field lattices having fewer grid points to be used without any significant loss of accuracy. The resulting processing speed of about 0.25 s per conformation on a 2.4 GHz Intel Pentium processor allows the method to be used for virtual screening on commonly available desktop machines. Moreover, the results indicate that grid-based superposition methods could be efficiently used for the virtual screening of compound libraries.

Keywords BRUTUS · Molecular superposition · Virtual screening

T. Rönkkö · A. J. Tervo · J. Parkkinen · A. Poso (✉)
Department of Pharmaceutical Chemistry, University of Kuopio, P.O. Box 1627, FIN-70211 Kuopio, Finland
e-mail: Antti.Poso@uku.fi

Present Address:
A. J. Tervo
AstraZeneca R&D Mölndal, Pepparedsleden 1, SC2, S-43183 Mölndal, Sweden

J. Parkkinen
Department of Computer Science, University of Joensuu, P.O. Box 111, FIN-80101 Joensuu, Finland

Introduction

Finding novel lead molecules is one of the primary goals of the early phases of drug discovery projects [1]. However, often there is no structural information available about the therapeutic target of interest [2–4], which limits the usefulness of receptor-based methods such as molecular docking [5]. In such situations, ligand-based methods are usually chosen. In particular, the superposition of ligands is the method of choice for preparing data for subsequent similarity analysis and virtual screening [2].

Over the years many molecular properties have been investigated and numerous algorithms presented to allow superimposing of molecules. For example, Krämer et al. [6] superimposed molecules according to the features derived from the structure, Cosgrove et al. [7] utilized the shapes of the molecular surface, Lemmen et al. [2] used partitioning and incremental construction, Kearsley and Smith [8] used atomic partial charges and steric volumes, while Mills et al. [9] superimposed molecules by using hydrogen-bond maps. All of these methods can be applied to the screening of compound databases, but each method is based on a different concept with different characteristics and subsequent limitations. As a result, one particular method may work better in some contexts than the others, but it may be difficult to predict which method will actually produce the desired results [1]. Thus, multiple search methods need to be used.

Molecular energy fields are suitable for elucidating structurally novel and biologically active molecules, since energy fields model molecular structures over 3D space surrounding atomic nuclei. This space around the atomic nuclei relates directly to non-covalent

interactions that are thought to be mainly responsible for the biological effects at the molecular level [10]. It is well known that structurally distinct compounds may exhibit similar biological activities, and, assuming that the mechanism of action is the same, the energy fields of dissimilar compounds should be similar. If so, a virtual screening method that finds similar energy fields should be able to come up with both structurally novel and biologically active compounds.

Molecular energy fields can be presented by rectilinear lattices and analytic Gaussian functions [11]. In the first approach, an energy field consists of a three-dimensional lattice where the energy between a probe atom and a molecule is computed at each grid point. In the latter approach, an energy field is formed by attaching a Gaussian function to every atom, and the union of these functions forms an energy field.

The similarity of two molecules in a given alignment can be evaluated using either of the methods, although the use of analytic Gaussian functions has been reported to increase the efficiency of similarity evaluation by as much as three orders of magnitude when compared to a grid-based evaluation of similarity [12]. However, we claim that the grid-based evaluation of similarity is still a viable approach, and it is possible to improve the efficiency of the grid-based evaluation to a more usable level by using coarse energy field lattices and a simple interpolation algorithm that reduces the inherent loss of accuracy without consuming too much computational resources.

In this paper, the theory underpinning the BRUTUS algorithm is discussed and the method is parameterized by molecular self-overlap studies. In our recent paper [13], the algorithm was validated by applying BRUTUS to both pair-wise alignments and lead finding.

Methodology

Hodgkin index

The similarity of energy fields A and B can be estimated by the Hodgkin [14] hodgkin index

$$H_{\mathrm{AB}} = \frac{2\int_V p_{\mathrm{A}} p_{\mathrm{B}} dV}{\int_V p_{\mathrm{A}}^2 dV + \int_V p_{\mathrm{B}}^2 dV} \tag{1}$$

where p_{A} and p_{B} are density functions of energy fields A and B [14].

With the Hodgkin index [14], the similarity of any two molecules can be easily estimated by keeping either of the molecules in a fixed position while the other is rotated and translated [15]. After each step, an energy field of the moving molecule is generated, and the similarity is estimated by the Hodgkin index. When enough of these steps are repeated, an optimum superposition can be obtained.

Although straightforward, the above procedure is not an effective way to estimate similarity during optimization of the Hodgkin index. The main problem with this simple approach is that energy fields must be recomputed after each optimization cycle in order to estimate the similarity of energy fields in the subsequent alignment. However, it is also possible to rotate the energy field instead of the molecule, in which case the energy field does not need to be recomputed during an optimization process that may involve thousands of similarity evaluations [16]. Rotating the energy fields instead of molecules is especially useful if the energy fields are computed by quantum mechanical methods [17].

In BRUTUS, energy fields are expressed as rigid rectilinear lattices, and a *coordinate transformation* is applied to map the coordinates between energy fields to achieve rotation and translation. After this transformation, the similarity of an alignment is estimated by the Hodgkin index such the grid points of fixed and moving energy are associated by another coordinate transformation, and the energy values of static energy field are approximated by an *interpolation* algorithm.

Coordinate transformation

A coordinate system of an energy field can be represented by vector t_a that identifies an origin of the coordinate system and vectors x_a, y_a and z_a, which are the basis of the X, Y and Z-axis. In such a system, vector q_i can be mapped from the internal coordinate system of a molecule to point q_w in a world coordinate system by coordinate transformation [18]

$$q_w = C(x_a, y_a, z_a, t_a) \cdot q_i \tag{2}$$

where

$$C(x_a, y_a, z_a, t_a) = \begin{bmatrix} x_{a_x} & y_{a_x} & z_{a_x} & t_{a_x} \\ x_{a_y} & y_{a_y} & z_{a_y} & t_{a_y} \\ x_{a_z} & y_{a_z} & z_{a_z} & t_{a_z} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Likewise, vector q_w, expressed in the world coordinate system, can be transformed back to an internal coordinate system defined by vectors x_a, y_a, z_a and t_a by inverse coordinate transformation

$$q_i = C^{-1}(x_a, y_a, z_a, t_a) \cdot q_w \tag{3}$$

Energy fields can be positioned and oriented using three-dimensional homogenous coordinate transformations, which are described by 4×4 matrices [18, 19]. The matrix for translating a coordinate system by vector v is [19]

$$T(v) = \begin{bmatrix} 1 & 0 & 0 & v_x \\ 0 & 1 & 0 & v_y \\ 0 & 0 & 1 & v_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

In 3D applications, objects can be rotated around any line in space, but the simplest rotation axes to work with are those that are parallel to the X, Y and Z coordinate axes [18]. By combining these primitive rotations in a different order, objects can be oriented in 3D space. There are many possible orderings of the rotations, and it is not necessary to use all three coordinate axes [20]. For example, all possible orientations of an object can be achieved by rotating first around the Z axis, then the Y and finally again around the Z axis. In a Euler angle representation, these axes of rotation are fixed to the object, and the rotations are expressed in a left to right order.

The matrix for rotating a vector around to the Y axis by angle β is [19]

$$R_y(\beta) = \begin{bmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5)$$

The Z-axis rotation matrix by angle γ is [19]

$$R_z(\gamma) = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 & 0 \\ \sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

The combined coordinate transformation for positioning and orienting a centered energy field is

$$M = T(v) \cdot R_z(\alpha) \cdot R_y(\beta) \cdot R_z(\gamma) \cdot C(x_a, y_a, z_a, t_a) \quad (7)$$

where $\cdot$ is the matrix multiplication. With Equation 7, the location of a grid point within the energy field lattice is transformed to a world coordinate.

Interpolation

As the energy fields are rotated, the grid points of the template and the database molecules will not exactly match. Thus, the energy between the grid points must be estimated by an interpolation method. Perkins et al. [16] solved this problem by using a simple interpolation scheme where lattices were overlaid and the closest grid point was used. However, the approximation error increased by the resolution, and, as a consequence, fairly dense 1.0 Å grids had to be used. De Cáceres et al. [17] solved the very same problem by using a distance-based interpolation function that considered many neighboring grid points. Thus, the resulting energy function was smooth and easy to optimize.

However, the interpolation is the most time consuming phase of the grid-based similarity evaluation, and considering multiple grid points can make the computation unwieldy. In order to speed up the similarity search even further, the interpolation process must be as simple as possible and coarse grids have to be used. Needless to say, the coarser the grid, the quicker the calculation of the similarity is, but this must be balanced by loss of accuracy [21].

In the simplest possible interpolation technique, the interpolated value is computed from one single grid point; i.e., the energy fields are superimposed and the energy of the grid point nearest to the point of interest is taken. Although simple, this method is crude and can create discontinuity if coarse grids are used. However, if either of the energy fields is expressed at a much higher resolution than the other, then the accuracy of the method can be increased.

Using this simple interpolation method, Equation 1 can be written as

$$H_{AB} = \frac{2u_a \sum_{k=1}^{d_a} \sum_{j=1}^{h_a} \sum_{i=1}^{w_a} A_{ijk} B_{i'j'k'}}{u_a \sum_{k=1}^{d_a} \sum_{j=1}^{h_a} \sum_{i=1}^{w_a} A_{ijk}^2 + u_b \sum_{k=1}^{d_b} \sum_{j=1}^{h_b} \sum_{i=1}^{w_b} B_{ijk}^2} \quad (8)$$

where A_{ijk} is a grid point at position (i,j,k) in an energy field lattice A. The grid point (i,j,k) in the energy field A is mapped to point (i',j',k') in energy field B by coordinate transformation

$$\begin{bmatrix} i' \\ j' \\ k' \\ 1 \end{bmatrix} = C^{-1}(x_b, y_b, z_b, t_b) \cdot M \cdot \begin{bmatrix} i \\ 1 \end{bmatrix} \quad (9)$$

The different resolutions of the energy fields are taken into account by u_a and u_b, which are the volumes of the unit cells of energy fields A and B. The volume of unit cell is computed from unit vectors that describe the coordinate system of the energy field as

$$u_a = x_a \times y_a \bullet z_a \quad (10)$$

where $\times$ is the cross product (vector product) and $\bullet$ is the dot product.

Figure 1 shows how the grid points are mapped from energy field A to energy field B using equation 9. Filled circles represent the grid points of energy field A, whereas open circles represent the grid points of energy field B. For example, grid point (1,1) is mapped to point (2,2) and grid point (2,1) is mapped to (4,2). If the grid separation of energy field B is increased, the precision of interpolation algorithm can improved, since the difference between the mapped grid points becomes shortened.

Systematic search

Assuming that a reasonable superposition for any two molecules exists, the real question is how to maximize the likelihood that it will be found. Gradient-based optimization methods, for example, can only go downhill on the energy surface and so they can only locate the minimum that is nearest (in a downhill sense) to the starting point [22]. A systematic search, on the other hand, is a reliable but very slow method to find global and local maxima [23].

Selecting the most representative set of starting positions is of primary importance in finding global maxima with Gradient optimization methods [15]. In theory, there is an infinite number of possible alignments to be considered but, in practice, not every alignment has to be evaluated. It is sufficient to select a small set of starting positions and optimize these with a Gradient-based optimization method to find near optimal solutions.

Grant et al. [24] found it convenient to center molecules and start with four initial alignments that were derived from principal axis while Mestres et al. [25] used 208 unique starting positions. However, while a small number of starting positions may be sufficient for aligning molecules that are approximately the same size, it is debatable whether the number of starting positions is adequate for partial matches where a small molecule is aligned onto a larger molecule, or vice versa. It would be logical to assume that the number of starting positions should increase with the size of a molecule, or at least the starting positions should be chosen with greater care, if partial matches are desired.

In BRUTUS, possible starting positions are examined by a systematic search, after which the most potential starting positions are optimized by a Gradient-based method. A systematic search of alignment space yields with a suitably large *rotation step* and a *translation step* a broad range of potential starting positions from which to choose, irrespective of the molecular size or initial alignment. By *optimizing* the very best alignments from this initial search phase, a set of interesting solutions can be produced.

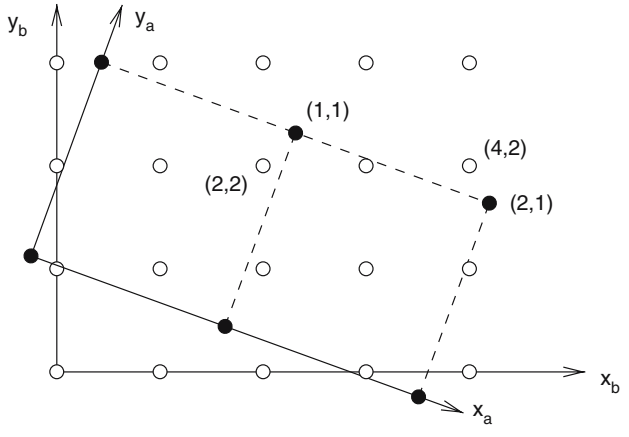

Fig. 1 Grid points (1,1) and (2,1) of energy field A (black circles) are mapped to grid points (2,2) and (4,2) of high-resolution energy field B (open circles)

Rotation step

An experiment was conducted to investigate the effect of rotation and to determine a suitable rotation step for the systematic search. For this purpose a set of 300 random molecules was selected from the Maybridge chemical database [26] prepared by Tervo et al. [13] The set contained 100 molecules with 20 atoms, 100 molecules with 40 atoms and 100 molecules with 60 atoms. The molecules were centered, and each molecule was superimposed with its rigid copy. The identical copies were gradually rotated from –90 to 90°, according to the X, Y and Z rotation axes, and the energy fields were computed for both the molecule and its identical copy with a cut-off value of 5 kcal/mol.

Figure 2 illustrates how the Hodgkin index changes when a perfectly overlaid pair of molecules is gradually rotated to separate the charge centers. Two conclusions can be drawn from the figure; (1) it may be possible to optimize alignment of energy fields by a Gradient-based optimization method if the optimal alignment is less extensive than $\pm$ 90°. After all, the average similarity coefficient is steadily increasing as the optimal alignment is approached. (2) As the molecular size increases, the similarity drops more sharply. This suggests that the number of initial alignments is dependent on the molecular size, and, a method that has been parameterized for small molecules, may not work for large molecules such as proteins.

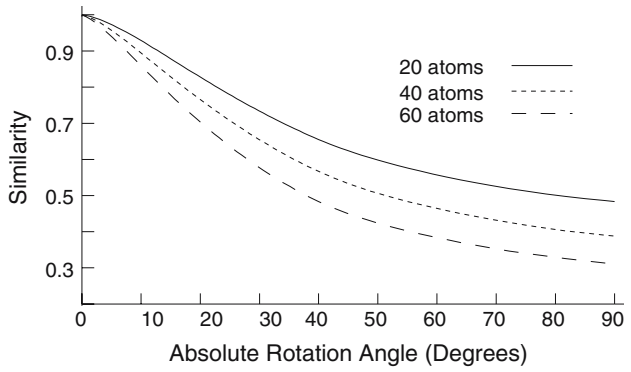

Fig. 2 The effect of rotation on the average Hodgkin index computed for self-overlays of 100 molecules containing 20, 40 and 60 atoms

In order to decide upon a suitable step size, the standard deviation in Fig. 3 was also taken into account. The figure shows the standard deviation and the average Hodgkin similarity index that were computed while gradually rotating self-overlays of 100 randomly selected molecules, containing 40 atoms, from –180 to 180°. These results allow one to draw the conclusion that alignments within ± 30° from the optimal alignment are clearly distinguishable from the majority of orientations. In this area, the average Hodgkin index is likely to be above 0.55, whereas outside that window, an average Hodgkin index is likely to be below 0.55. It is these starting positions within ± 30° from the optimal alignment that can be reliably optimized to reach the optimum.

By lowering the threshold, the step size can be increased, thus improving the speed of the systematic search. However, at the same time the probability of finding the global maximum will decrease. If a starting position for Gradient-based optimization is too distant from the global maximum, the optimization may instead advance towards local maximum. Gradient-based optimization methods usually advance in the direction of increasing similarity, and whenever two atoms come close to each other, there is a potential local maximum. Therefore, the starting position has to be relatively close to the global maximum if it is to be identified reliably by Gradient-based optimization methods. In order to ensure that at least one starting position is close enough to the global maximum, possible starting positions should be explored with a step size, and, in order to locate partial matches, the alignment space should be explored uniformly without disregarding some areas.

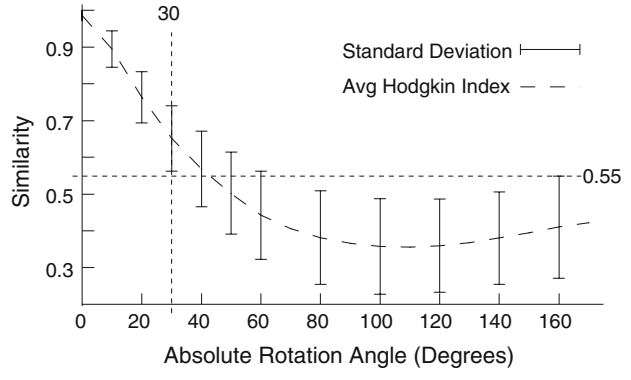

Fig. 3 Average Hodgkin index and standard deviation of 100 gradually rotated self-overlays containing molecules with 40 atoms

Translation step

A similar study was conducted to estimate the step size of the translation. The results in Fig. 4 show the effect of translation on an average Hodgkin index, computed for self-overlays of 200 molecules, which contain 100 randomly selected molecules with 20 atoms and 100 randomly selected molecules with 60 atoms. Here, the average Hodgkin index decreased steadily, but the minimum and the maximum similarity values shown by the dashed lines do not follow the average trend. For example, at point A, the similarity after translating a particular molecule by just 2.0 Å is as low as 0.2. At point B, the similarity is above 0.3, even though the molecule in question has been translated by more than 10 Å.

These two extreme cases are explained by examining the results from planar molecules in the test set. With planar molecules, the translation may have a dramatic effect on the similarity when two perfectly overlaid molecules are translated to separate the layers. If, on the other hand, a planar molecule is

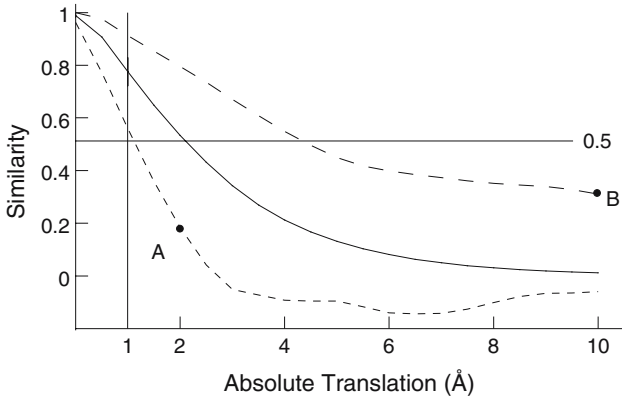

Fig. 4 The effect of translation on an average Hodgkin index (solid line) computed for self-overlays of 200 randomly selected molecules. The dotted and the dashed lines show the minimum and the maximum similarity, respectively

translated in a plane, the similarity is not likely to change greatly. Both cases need to considered when estimating the number of trial alternatives to be investigated.

Figure 4 shows that in order to find alternative starting positions, which may be further optimized, the distance between the optimal alignment and the starting position cannot be much more than ± 1.0 Å if results are desired with a high degree of confidence. The Hodgkin index within the ± 1.0 Å window is clearly above 0.5, whereas the Hodgkin index drops below 0.5 even for the most troublesome alignments. However, if the starting positions are explored by using a larger step size, the Hodgkin index of the most potential starting positions cannot be reliably distinguished from the Hodgkin index of the poor positions, thus increasing the probability of the algorithm to choose a sub-optimal starting position and the optimizer to stick on a local maximum.

Optimization

The Hodgkin index is optimized according to six degrees of freedom (three rotations and three translations), and the rotation angles α and β are made to vary between 0 and 360° whereas the γ angles range from 0 to 180°. With the step size previously established, it should therefore be possible to find interesting starting positions by substituting 30, 90, 150, 210, 270 and 330° for α and β, and 30, 90 and 150° for γ in the Equation 7. Thus, for each trial translation, only 108 rotations ($6 \times 6 \times 3$) need to be evaluated in order to find at least one starting position within ± 30° from the optimal alignment.

If one wishes to optimize the three translational parameters of Equation 1, it is assumed that the center point of the template molecule is located inside a bounding box that is defined by the outermost heavy atoms of the database molecule. The volume of this bounding box defines the number of trial positions that must be evaluated to find the starting positions for further optimization. Using the ± 1.0 Å window, an average of 80 trial alignments are needed for a typical molecule containing 40 atoms. The number of trial alignments naturally depends on the size of the molecule, which in turn depends on the number of atoms.

Considering the optimization of both the rotational and translational variables, some 8500 trial alignments are needed to find 256 starting positions for further optimization. Furthermore, some 3500 more trial alignments are needed to optimize these starting positions. Thus, about 12,000 trial alignments need to be evaluated for a typical molecule containing 40 atoms. Compared to algorithms using fewer starting positions, BRUTUS is a brute-force algorithm that attacks the problem of finding pair-wise alignments in a rather brutal way, hence the name.

Results and discussion

Euler angles and quaternions

Rotations can be achieved by Euler angles or quaternions [20]. We found Euler angles and transformation matrices convenient for use in BRUTUS, as a number of translation, rotation and scaling operations could be encapsulated in a single 4×4 matrix. With quaternions, these operations are handled separately, and a scalar multiplier and translation vector would have to be operated along with the transformation matrix to mimic the homogeneous coordinate transformation of BRUTUS. However, quaternions might improve the efficiency of minimization [24, 27], and it might be useful to replace rotations of Equation 7 with a quaternion formulation.

The problem with our current approach is that rotation $R_z(\alpha)\cdot R_y(\beta)\cdot R_z(\gamma)$ has a singularity when β is close to zero; i.e., change of α can undo change in γ when β approaches zero. This gives rise to artificial saddle points and maxima, which may affect the optimization of the molecular overlay[24]. However, in this study the artificial saddle points and maxima did not represent a major problem.

The number of starting positions needed

Selecting the most representative set of starting positions is essential if one is to find the global maximum with Gradient-based optimization methods [15]. In BRUTUS, potential starting positions are identified by translating and rotating molecules systematically with a step size of 2 Å and 60°, after which at most 256 of the best alignments are used as starting positions in the subsequent optimization. Due to these fairly small steps, some 8500 trial alignments must be evaluated before the most potential starting positions can be selected.

It can be argued that a more elaborate optimization scheme would cope with far fever trial alignments. However, we believe that a large number of alignments has to be evaluated in order to find decent starting positions for finding partial matches, where a small molecule is aligned to a larger molecule, or vice versa. Database searches are often conducted to find compounds that are structurally novel [1], and molecules

that are larger or smaller than a template molecule may also be interesting in that respect. Moreover, it should be noted that also the more elaborate optimization algorithms, such as simulated annealing and evolutionary algorithms, require a set of starting positions (initial population) in order to locate the optimal alignment. If the potential starting positions cannot be identified by using a step size much greater than 2.0 Å and 60°, then it would be anticipated that, irrespective of the optimization scheme, the number of trial alignments cannot be greatly reduced if the potential starting positions and interesting local maxima are to be identified reliably.

It may also be useful to remember that the step sizes were derived from self-overlays, where an optimal superposition always exists. This is usually not the case with the structurally dissimilar molecules, which are likely to be encountered in virtual screening. However, if one wishes to locate reasonable starting positions for finding partial matches, it would be unrealistic to expect that interesting starting positions would be found reliably with a much larger step size in a more challenging scenario. Instead, in the view of the fact that two dissimilar structures do not usually have a perfect alignment, the step size of a systematic search may even need to be tightened to find the most potential starting positions in a real world scenario. Alternatively, the energy fields or the similarity function should be modified such that the potential starting positions could be identified with a much larger step size.

Precision of interpolation

In order to investigate the accuracy of Equation 8, a random set of 200 molecules was selected and superimposed with their identical copies. The molecules were centered, and one copy was kept in a fixed position while the other copy was rotated. For each rotation, the Hodgkin index was computed precisely and approximated by two different interpolation schemes. The precise similarity coefficients were obtained by rotating a molecule and re-generating the electrostatic energy field of the rotated molecule so that the grid points of the two fields matched perfectly. A resolution of 0.2 Å was used for both energy fields. The Hodgkin index was approximated by linear interpolation and the simple interpolation scheme while using two different resolutions. For the linear interpolation, both energy fields were generated with a resolution of 2.0 Å, and the eight nearest grid points were used to approximate the energy between the grid points. For the simple interpolation scheme, the template molecule was expressed with resolutions 2.0 Å and 1.0 Å in turn while the database molecule was expressed with 0.5 Å and 0.25 Å resolutions.

The correlation coefficient (R^2) was 0.899 by using linear interpolation. When the simple interpolation scheme was used with resolutions 2.0 Å and 0.5 Å, the correlation coefficient increased to 0.966, and, when using resolutions 1.0 Å and 0.25 Å for template and database molecules, the correlation coefficient was further increased to 0.999.

The results show that this computationally lightweight and simple interpolation scheme can accurately estimate similarity using coarse grids, even though the grid points of the template and database molecules do not coincide. However, due to the approximation error, some noisiness is still expected in the optimization of equation 8, and optimization algorithms that can cope with noisy data are preferred.

Energy fields

In principle, any type of energy field can be used with BRUTUS. However, energy fields with continuous, smoothly varying densities and large regions of positive, neutral and negative charge are preferred in order to minimize the approximation error and to allow starting positions to be explored using a fairly large step size. Energy fields with sudden changes of density and small regions of positive, neutral or negative charge are likely to incur greater approximation errors and may even require tightening of the steps of the systematic search in order to locate appropriate starting positions for further optimization. If new energy field types are used with BRUTUS, the level of the approximation error and the applicability of the 60 degree rotational step and 2.0 Å translational step needs to be verified.

Electrostatic and van der Waals energy fields are suitable for BRUTUS, as these fields contain large regions of positive, neutral and negative charges. However, both fields also contain rapid changes of energy, and the usability of these fields can be improved by removing or smoothening these rapid changes. In van der Waals energy field, the energy changes rapidly near to the van der Waals radius. Smoothening the transition of energy near to van der Waals radius both reduces the approximation error and makes the similarity function continuous thereby allowing easier optimization of the similarity. In electrostatic energy field, the energy changes rapidly near the atomic centers. However, the charge inside the van der Waals radius is not meaningful for estimation of similarity, and the rapid changes can be eliminated by assigning a fixed value to the grid points that fall inside the van der Waals radius [28].

In BRUTUS, electrostatic charge and van der Waals volume are combined into a single energy field. The van der Waals component allows molecules to be superimposed by their shape and the electrostatic component allows charged regions of molecules to be matched. The combined energy field is created such that the charge inside the van der Waals radius is set to a positive value, and the electrostatic charge is used outside the van der Waals radius. Furthermore, the transition between the van der Waals and electrostatic regime is smoothened.

In this scheme, the relative importance on shape and electrostatic fit can be controlled by selecting the appropriate value that is assigned to the grid points inside the van der Waals radius. A large value favors a shape match while a small value favors an electrostatic fit. In BRUTUS, the shape match is favored over the electrostatic fit, and a value of five is assigned to grid points inside the van der Waals radius. However, the importance of shape and electrostatic fit may depend on the target [29], and this value might not be optimal for all targets.

Evaluation of similarity using this combined energy field is twice as fast as can be achieved using two distinct energy fields. However, the relative importance of shape and electrostatic fit might be easier to control if two distinct energy fields were used. Moreover, it is not possible to combine more than two types of energy fields. For example, electrostatic and hydrophobic energy fields cannot both occupy the area outside the van der Waals radius in the combined energy field. If these energy field types are to be used, there should be a distinct energy field for each energy field type. The use of distinct fields could be especially useful in aligning compounds for comparative molecular field analysis (CoMFA [30]), where proper compound alignments are crucial for the quality and reliability of the model and superposition time is typically not an issue.

Accuracy of molecular alignments

The accuracy of molecular alignments was examined in a self-overlay experiment, in which 35 various-sized compounds, extracted from X-ray structures of protein–ligand complexes, were aligned with their identical copies. Initially, 1000 arbitrarily rotated and positioned orientations were generated for each compound. Next, these random orientations were superimposed with the original compound. Molecular fields were derived from Gasteiger–Hückel charges, using field resolutions of 1.0 Å and 2.0 Å in turn. Finally, heavy atoms root mean square deviation (RMSD) values were calculated between the template compound and each resulting superposition. Average RMSD values of the superpositions for each compound were calculated to obtain an estimate of the structural accuracy of the superpositions.

BRUTUS omitted 38 and 39 (0.1%) out of 35,000 superimposed orientations, and generated 39 (0.1%) and 104 (0.3%) reverse alignments for symmetrical compounds, when using field resolutions of 1.0 Å and 2.0 Å, respectively. These alignments were left out of the average RMSD values. The remaining average RMSD values for the 35 investigated compounds are presented in Table 1.

The average inaccuracy of the structural superposition for the field-based self-overlays of all 35 compounds was 0.28 Å and 0.55 Å for field resolutions of

Table 1 The average accuracy of 35 self-overlays using two resolutions

Compd	PDB ID[a]	Atoms[b]	RMSD (Å)	
			1.0 Å	2.0 Å
1	3pgh	31	0.284	0.593
2	6cox	37	0.251	0.496
3	2hwb	39	0.272	0.569
4	1r09	41	0.306	0.659
5	1ruc	45	0.288	0.600
6	5tln	46	0.252	0.470
7	1ele	48	0.247	0.501
8	7est	51	0.245	0.498
9	1rud	54	0.638	1.172
10	1a85	55	0.250	0.496
11	2rs3	57	0.745	1.409
12	1hwr	60	0.252	0.461
13	1mmb	63	0.258	0.483
14	5tmn	65	0.245	0.501
15	1tmn	68	0.262	0.544
16	1hpv	70	0.245	0.447
17	1bma	73	0.246	0.530
18	1dmp	76	0.239	0.472
19	1qbu	79	0.244	0.480
20	1b0e	81	0.255	0.492
21	1d4h	83	0.248	0.468
22	1g35	87	0.240	0.471
23	1ebz	89	0.238	0.464
24	1d4l	91	0.246	0.487
25	1qbr	92	0.237	0.445
26	1ebw	96	0.237	0.438
27	1hxw	98	0.245	0.488
28	1hef	101	0.245	0.476
29	1hos	105	0.240	0.479
30	2bpy	106	0.245	0.494
31	1ec2	108	0.241	0.461
32	1dif	116	0.511	0.716
33	1ody	116	0.240	0.593
34	1a8g	118	0.255	0.534
35	1hiv	120	0.240	0.474

[a] PDB ID of the X-ray structure where the cocrystallized ligand was obtained

[b] Number of atoms in the compound

1.0 Å and 2.0 Å, respectively. There were no major differences between the average RMSD values of the different compounds, except for compounds 9, 11 and 32. For these nearly symmetrical compounds, BRUTUS generated some reverse-oriented alignments, which increased the average RMSD.

The results show that BRUTUS can generate self-overlays with an acceptable accuracy irrespective of the starting orientation stored in the database. Moreover, the results suggest that the average accuracy of superposition of two different molecules is at least 0.55 Å. If greater accuracy is required, BRUTUS is recommended to be used with a field resolution of 1.0 Å.

Running time

In our recent article [13], three database searches were conducted with BRUTUS version 0.6.3. As these searches used three distinct template molecules with 37, 92 and 118 atoms, the results of these searches provide a hint on the performance of BRUTUS in a real world scenario. Table 2 summarizes the average superpositions times per conformation for these three database searches.

The results lead to the conclusion that the superposition time depends on both the template molecule and the database. If the size of the template molecule is increased from 37 to 162 atoms, the running time of the database search does not increase by the same extent. In fact, it seems that the actual running time of BRUTUS is dependent more on the size of the average database molecule rather than on the template molecule. This is to be expected as the bounding box of the database molecule directly affects the search region and the number of starting positions that are to be evaluated. The size of the template molecule, on the other hand, only affects the running time indirectly. A large template molecule has more grid points than a small molecule, but, unless these grid points are matched with a database molecule, the points have no effect on the running time.

Table 2 The average superposition times of three database searches

Template[a]	Atoms[b]	Time (s)
6cox	37	0.21
1qbr	92	0.24
1a8g	118	0.25

[a] PDB ID of the X-ray structure where the template molecule was extracted

[b] Number of atoms in the template molecule

Efficiency of grid-based algorithms

Efficiency is one of the most crucial features when deciding whether a method can be applied to real-world virtual screening problems that involve millions of molecules [1]. In two virtual screening experiments [13], BRUTUS has been shown to superimpose a conformation in less than 0.25 s, and this suggests that performance of BRUTUS is sufficient for practical applications. Moreover, the efficiency of BRUTUS indicates that grid-based methods in general can be efficient for comparing the similarity of molecular energy fields.

To improve the efficiency of the similarity calculations, in 1992 Good et al. [11] suggested that the effect of the atomic partial charges should be approximated by overlapping Gaussian functions. According to Good and Richards [12], Gaussian functions increase the efficiency of similarity evaluation by as much as three orders of magnitude compared to a grid-based evaluation of the similarity. However, Good and Richards utilized a 0.2 Å grid separation in their comparative study whereas BRUTUS uses a 2.0 Å separation. Moreover, BRUTUS applies an interpolation algorithm to avoid regenerating energy fields during similarity optimization. These changes make BRUTUS a different algorithm, and Good s conclusions may not apply to BRUTUS. For example, the mere change of grid separation from 0.2 Å to 2.0 Å represents a thousand-fold reduction in the number of grid points, and this can compensate for the thousand-fold increase in performance that Good and Richards [12] achieved by using Gaussian functions instead of grid-based evaluation. It is therefore possible that a grid-based similarity calculation method coupled with an interpolation algorithm can be as efficient as a Gaussian function similarity calculation.

Conclusions

BRUTUS is an automated computer program for rigid-body molecular superposition. BRUTUS identifies potential starting positions for further optimization via a systematic search, and, during a single molecular superposition, some 8500 alignments are evaluated to find up to 256 most potential starting positions. Another 3500 alignments are then evaluated to optimize these starting positions. This seemingly large number of trial alignments makes BRUTUS reliable, and interesting alignments are located independently of the starting position.

BRUTUS is a molecular field-based method, in which coordinate systems are transformed to estimate the similarity of energy fields in various alignments. This removes the need to re-generate energy fields during optimization, and thereby makes the algorithm more efficient. Moreover, BRUTUS applies a simple interpolation algorithm for estimating the energy between grid points, and this allows coarse energy fields to be used without significant loss of accuracy. The resulting superposition speed of about 0.25 s per conformation on a 2.4 GHz Intel Pentium processor allows the algorithm to be used for virtual screening of large molecular databases on commonly available desktop machines.

The results indicate that grid-based superposition methods can be efficiently used for the virtual screening of compound libraries. Moreover, these results open new possibilities for constructing virtual screening algorithms and for finding structurally diverse lead molecules for a variety of targets.

Acknowledgments The authors gratefully acknowledge the National Technology Agency of Finland (TEKES) and Finncovery Ltd. for their financial support. Orion Pharma Ltd., Juvantia Pharma, Visipoint Ltd. and CSC-Scientific Computing Ltd. are acknowledged for their help in the software development. The authors also thank Petri Kokkonen, Outi Salo and Juha Haataja for their comments and healthy criticism.

References

1. Sheridan RP, Kearsley SK (2002) Drug Discov Today 7:903
2. Lemmen C, Lengauer T, Klebe G (1998) J Med Chem 41:4502
3. Labute P, Williams C, Feher M, Sourial E, Schmidt JM (2001) J Med Chem 44:1483
4. Melani F, Gratteri P, Adamo M, Bonaccini C (2003) J Med Chem 46:1359
5. Lyne PD (2002) Drug Discov Today 7:1047
6. Krämer A, Horn HW, Rice JE (2003) J Comput Aid Mol Des 14:13
7. Cosgrove DA, Bayada DM, Johnson AP (2000) J Comput Aid Mol Des 14:573
8. Kearsley SK, Smith GM (1992) Tetrahedron Comput Methodol 3:615
9. Mills JEJ, Perkins TDJ, Dean PM (1997) J Comput Aid Mol Des 11:229
10. Parretti MF, Kroemer RT, Rothman JH, Richards WG (1997) J Comput Chem 18:1344
11. Good AC, Hodgkin EE, Richards WG (1992) J Chem Inf Comput Sci 32:188
12. Good AC, Richards WG (1993) J Chem Inf Comput Sci 33:112
13. Tervo AJ, Rönkkö T, Nyrönen TH, Poso A (2005) J Med Chem 48:4076
14. Hodgkin EE, Richards WG (1987) Int J Quantum Chem 14:105
15. Manaut F, Sanz F, Josè J, Milesi M (1991) J Comput Aid Mol Des 5:371
16. Perkins TDJ, Mills JEJ, Dean PM (1995) J Comput Aid Mol Des 9:479
17. De Cáceres M, Villà J, Lozano J, Sanz F (2000) Bioinformatics 16:568
18. Hearn D, Baker MP (1997) Computer graphics: C version. Prentice Hall, New Jersey, pp. 407–423
19. Foley JD, van Dam A, Feiner SK, Hughes JF, Philip RL (1997) Introduction to computer graphics. Addison-Wesley, Reading, pp 171–191
20. Parent R (2002) Computer animation: algorithms and techniques. Morgan Kaufmann, San Francisco, pp 51–57
21. Burt C, Richards WG, Huxley P (1990) J Comput Chem 11:1139
22. Leach AR (2001) Molecular modelling: principles and applications, 2nd edn. Pearson Education, Harlow, pp 255–257
23. Turner DB, Willett P, Ferguson AM, Heritage TW (1995) SAR QSAR Environ Res 3:101
24. Grant AJ, Gallardo MA, Pickup BT (1996) J Comput Chem 17:1653
25. Mestres J, Rohrer DC, Maggiora GM (1997) J Comput Chem 18:934
26. Maybridge Chemicals Database, Maybridge Chemicals Company Ltd., Cornwall, U.K.
27. Masek BB, Merchant A, Matthew JB (1993) J Med Chem 36:1230
28. Good AC (1992) J Mol Graphics 10:144
29. Barnaby M, Gutiérrez-de-Terán H, Sanz F, Villà-Freixa J (2004) Proteins 56:585
30. Cramer III RD, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959