JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Enhancing Molecular Shape Comparison by Weighted Gaussian Functions

Xin Yan,[†] Jiabo Li,*[,‡,§] Zhihong Liu,[†] Minghao Zheng,[†] Hu Ge,[†] and Jun Xu*[,†]

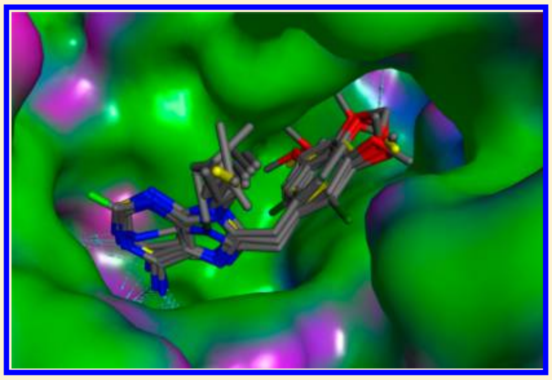
[†]Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-sen University, 132 East Circle at University City, Guangzhou, 510006, China

[‡]SciNet Technologies, 9943 Fieldthron Street, San Diego, California 92127, United States

[§]Accelrys, Inc., 10188 Telesis Court, San Diego, California 92121, United States

Supporting Information

ABSTRACT: Shape comparing technologies based on Gaussian functions have been widely used in virtual screening of drug discovery. For efficiency, most of them adopt the First Order Gaussian Approximation (FOGA), in which the shape density of a molecule is represented as a simple sum of all individual atomic shape densities. In the current work, the effectiveness and error in shape similarity calculated by such an approximation are carefully analyzed. A new approach, which is called the Weighted Gaussian Algorithm (WEGA), is proposed to improve the accuracy of the first order approximation. The new approach significantly improves the accuracy of molecular volumes and reduces the error of shape similarity calculations by 37% using the hard-sphere model as the reference. The new algorithm also keeps the simplicity and efficiency of the FOGA. A program based on the new method has been implemented for molecular overlay and shape-based virtual screening. With improved accuracy for shape similarity scores, the new algorithm also improves virtual screening results, particularly when a shape-feature combo scoring function is used.



INTRODUCTION

Molecular shape is a very important concept in computer-aided drug design. High throughput virtual screening based on molecular shape similarity has been widely used in drug discovery.[1] Over the past two decades, various methods for comparing molecular shapes have been developed[2−13] and numerous applications for virtual screening, scaffold hopping, and shape-feature based molecular alignment have been made in the past.[8−17] New drug discovery paradigms, such as network pharmacology or system pharmacology, demand faster and more effective methods for lead identification and optimization.[18,19] Generally speaking, the molecular shape comparing methods can be classified into two major categories:[20] those involving explicit shape comparison[2−4,11−13] and those based on shape-descriptor comparison.[5−7] The descriptor based shape comparison can be extremely fast. For example, one of the recently developed methods is Ultrafast Shape Recognition (USR) by Ballester and Richards,[5] and it can be used to screen millions of shapes in a second by using 12 shape descriptors. However, the representation of a shape using finite number of descriptors can never be completed; therefore, the descriptor based similarity comparison is generally considered as less effective than the explicit shape methods. A recent study by Nicholls et al.[1] shows that the shape similarity computed with various descriptor-based methods correlates poorly with shape-Tanimoto similarity computed with ROCS (Rapid Overlay of Chemical Structures), the current gold standard for shape comparison. The incomplete description of a shape by limited number of descriptors makes these methods "fragile".[1] For example, in some cases, two similar shapes can be seen as quite different via a descriptor-based method (i.e., a false negative). These false negatives are unacceptable when a virtual screening tool is used as a fast compound filter. The explicit shape comparison methods are usually considered to be more effective since the best alignments for shape matching are generally obtained in these approaches, and some useful information can be detected as to explain why some ligands are more active than the others. However, the explicit shape comparison method is usually less efficient. Perhaps the most practical approach is the combination of both types of methods. Actually, such a hybrid approach has been implemented in a widely used pharmacophore modeling software suite Catalyst,[21] which is now fully integrated in Discovery Studio of Accelrys.[22] For a shape-based 3D database search in Catalyst, a prefilter based on molecular volume and moments of inertia is used to prescreen molecules, and only the molecules with potential shape similarity (i.e., they passed the prescreen) are further examined by direct shape comparison using grids.[21]

There are many methods of representing molecular shapes,[20] the two most widely adopted methods are based on the hard-sphere[13,23] and Gaussian-sphere models.[2,3] Even though

Received: December 13, 2012
Published: July 11, 2013

dx.doi.org/10.1021/ci300601q | *J. Chem. Inf. Model.* 2013, 53, 1967−1978

analytical expression for the volume and its derivatives are available for the hard-sphere model,[24,25] the implementation is not entirely trivial because the formulas for intersection of multiple spheres become increasingly complex. Another issue that prevents efficient optimization is the discontinuity of derivatives of overlapping volumes.[13] Great simplification is achieved when the hard spheres are replaced by Gaussian spheres.[2,3,11,12] The use of Gaussian functions for comparing molecular similarity[11,12] has a deep connection with the ideas introduced to quantum chemistry by Boys,[26] in which atomic orbitals are represented by linear combination of Gaussian functions. Good and Richards[11,12] used Gaussian functions to fit the electron densities of atomic orbitals, and an analytical method was derived for computing shape similarity. However, the Gaussian electron-density based shape similarity was not very accurate compared to the hard-sphere model of atoms with VDW radii. The major progress for Gaussian function approach was made by Grant and Pickup.[2,3] In their seminal papers, a highly accurate representation of molecular shape based on overlapping Gaussian spheres was given. A molecular volume is represented as the summation of alternative inclusion−exclusion intersections with each intersection expressed as the integral of a set of overlapping Gaussian spheres. According to Grant and Pickup, high accuracy (error < 0.1%) can be achieved for molecular volume when up to sixth order intersections are included in the summation. Analytical expressions for each term and its derivates with respect to atom coordinates were given. Perhaps the most widely used implementation of the Gaussian function method is ROCS.[19] It is claimed that ROCS can calculate several thousand volumes per second while calculating intersections up to the sixth order. However, ROCS has also introduced a number of short-cuts for efficiency for computing overlap volumes between molecules.[4] For instance, all hydrogen atoms are ignored as they make very small contribution for the overall molecular shape, and all heavy atoms are set with equal radii. The most critical simplification in ROCS is that the shape density function of each molecule contains only the first order terms, and all higher order terms in the original Gaussian approach[2,3] are omitted. (Note: in ref 4, such an approach was called zero order Gaussian approximation, but we think it is better to call this approach the First Order Gaussian Approximation (FOGA) since the shape density includes only the first order Gaussian functions and all higher order terms consisting of the products of multiple Gaussian functions are ignored.) This significantly simplified computations but also received some criticism for the inaccuracy of this approximation.[5] One obvious issue is that the molecular volumes are significantly overestimated even though part of the error can be canceled in Tanimoto similarity calculation since both numerator and denominator are overestimated. Since the Gaussian shape algorithms are widely used in various virtual screening methods, it is important to investigate the errors introduced to the shape similarity calculation due to this overestimation of the volumes and to find a feasible solution to improve the accuracy in shape similarity calculations.

In this paper, we introduced a simple modification of the first order approximation. In this method, the molecular shape density is represented as the linear combination of weighted atomic Gaussian functions, and the weight for each atom is a simple function of the sum of the overlap between this atom and all others. Performance validation studies are reported. The rest of the article is organized as follows. A brief review of the Gaussian method is given, the potential problem of the first order Gaussian approximation is discussed, and a modified method of the first order Gaussian approximation is subsequently introduced. The accuracy of the new method is then studied, and simple case studies are reported.

■ METHODS

Hard-Sphere Method and Gaussian Method. For the clarity of discussion, some of the fundamental mathematical notations are introduced here, and most of the notations follow Grant and Pickup's original papers.[2,3] Shape and volume are two closely related concepts. A shape can be mathematically represented by a shape density function or by characteristic functions,[2,3] and the corresponding volume can be defined as the integral of the function over the three-dimensional space. The simplest description of a molecular shape can be considered as a set of fused hard-spheres, where each sphere presents an atom with its van der Waals radius. In a hard-sphere model, the shape density function $H(r)$ takes a simple value, either 1 if the coordinate r is within the molecule or 0 if r is outside. The shape density function $H(r)$ can be expressed in terms of the shape density functions of individual atoms and their overlaps:

$$H(r) = \sum_{i} h_i(r) - \sum_{i<j} h_i(r)h_j(r) + \sum_{i<j<k} h_i(r)h_j(r)h_k(r) - \sum_{i<j<k<l} h_i(r)h_j(r)h_k(r)h_l(r) + ... \tag{1}$$

The first summation runs over all atoms, and all atom overlaps should be subtracted as in the second summation over all atom pairs. All three atom overlaps should be added back, and the alternative inclusion and exclusion terms go on up to the order of n (n is the number of atoms in a molecule). For a hard sphere model, $h_i(r)$ can be expressed as the following:

$$h_i(r) = \begin{cases} 1, & (|r - r_i| \le \sigma_i) \\ 0, & (|r - r_i| > \sigma_i) \end{cases} \tag{2}$$

where σ_i is the van der Waals radius of atom i and r_i is the position of atom i.

Since $H(r)$ has a value either 1 or 0, one obvious property of $H(r)$ in the hard-sphere model is shown as follows:

$$H(r)^2 = H(r) \tag{3}$$

According to eq 1, the volume of a molecule can be calculated as

$$V = \int H(r)\, \mathrm{d}r = \sum_{i} v_i - \sum_{i<j} v_{ij} + \sum_{i<j<k} v_{ijk} - \sum_{i<j<k<l} v_{ijkl} + ... \tag{4}$$

where v_i is the volume of atom i, which is

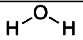
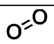
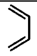
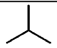
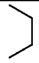
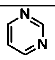
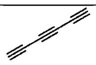
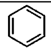
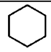
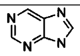
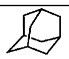
$$v_i = \frac{4\pi\sigma_i^3}{3} \tag{5}$$

and the intersection volumes of atoms pairs is defined as

$$v_{ij} = \int h_i(r)h_j(r)\, \mathrm{d}r \tag{6}$$

The higher order intersections are defined as integrals of the productions of multiple atoms' density functions.

Table 1. Overestimation Ratio of Self-Overlapping Volumes by First Order Gaussian Approximation (FOGA) As Compared to the Hard-Sphere Model

Molecule	Structure	Hard-Sphere Molecular volume	Self-overlapping volume by FOGA	Overestimation Ratio
Water (H_2O)		21.161	56.445	2.667
Oxygen (O_2)		25.949	52.866	2.037
Argon (Ar)	Ar	27.833	27.833	1.000
Methane (CH_4)	CH4	30.086	110.522	3.674
Formaldehyde (CH_2=O)		34.687	107.340	3.095
Hydrogen Cyanide (HCN)		37.987	102.800	2.706
Acetylene (C_2H_2)		39.075	120.365	3.080
Ethylene (C_2H_4)		43.019	165.430	3.846
Cyclopropane (C_3H_6)		58.495	276.029	4.719
Butadiene (C_4H_6)		72.700	324.238	4.460
Isobutane (C_4H_{10})		82.195	413.368	5.029
Butane (C_4H_{10})		82.302	410.689	4.990
Pyrimidine		86.384	448.905	5.197
C_6H_2		89.466	339.449	3.794
Benzene (C_6H_6)		90.267	475.376	5.266
Cyclohexane (C_6H_{12})		105.312	604.793	5.743
Purine		121.411	710.346	5.851
Adamantane		152.178	999.080	6.565

In the original Gaussian method,[2,3] the shape density of a molecule is defined in a way that is very similar to the approach above, except that the hard-sphere density functions of atoms are replaced by Gaussian functions:

$$G(r) = \sum_i g_i(r) - \sum_{i<j} g_i(r)g_j(r) + \sum_{i<j<k} g_i(r)g_j(r)g_k(r) - \sum_{i<j<k<l} g_i(r)g_j(r)g_k(r)g_l(r) + ... = 1 - \prod_i [1 - g_i(r)] \quad (7)$$

where

$$g_i(r) = pe^{-\left(\frac{3p\pi^{1/2}}{4\sigma_i^3}\right)^{2/3}(r-r_i)^2} \quad (8)$$

and p is an adjustable parameter controlling the softness of the Gaussian spheres. In the original Gaussian papers, Grant and Pickup showed that by setting $p = 2.7$, the molecule volumes can be calculated with accuracy within 0.1% by the following equation:

$$V^g = \int G(r)\,\mathrm{d}r = \sum_i v_i^g - \sum_{i<j} v_{ij}^g + \sum_{i<j<k} v_{ijk}^g - \sum_{i<j<k<l} v_{ijkl}^g + ... \quad (9)$$

where v_i^g is the volume of atom i, and the atom pair intersection v_{ij}^g can be calculated by

$$v_{ij}^g = \int g_i(r)g_j(r)\,\mathrm{d}r \quad (10)$$

and similarly for other higher order terms up to the sixth order. In this work, the originally suggested value[2] of 2.70 for parameter p is not used. Instead, p is set to $2\sqrt{2}$ (2.828427), so that the self-overlapping volume of a Gaussian sphere is equal to the volume of itself (as would be the case in a hard-sphere model). One nice property of Gaussian functions is that the product of two Gaussian functions is another Gaussian function, and the integral and derivatives can be easily calculated. The overlap volume of two molecules A and B is defined as the integral of the product of the Gaussian shape densities of two molecules:

$$V_{AB}^{g} = \int G_A(r)G_B(r)\,dr = \sum_{i\in A, j\in B} v_{ij}^{g} - \sum_{i,j\in A, k\in B} v_{ijk}^{g} - \sum_{i\in A, j<k\in B} v_{ijk}^{g} + \sum_{i<j\in A, k<l\in B} v_{ijkl}^{g} - .... \quad (11)$$

and the shape Tanimoto similarity can be defined as

$$S_{AB} = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}} \quad (12)$$

where V_{AA} is the self-overlap volume of molecule A and V_{BB} for molecule B. It can be proved that the similarity will be in the range of 0−1, and the similarity will be 1 only when two shape densities are identical.

First Order Gaussian Approximation (FOGA). According to eq 11, the overlap volume for two molecules is the summation of the contributions from all atom pairs, atom triplets, and higher order atom overlaps of Gaussian functions from both molecules. The summation includes all atom combinations, which leads to a combinatorial problem. To solve this problem in the original paper, only the overlaps from a group of neighbor atoms are considered, and this can be done by using atom neighbor lists. All other terms have little contribution and thus are ignored. As shown by Grant and Pickup,[2,3] only up to the *sixth* order of Gaussian overlaps are required to achieve high accuracy. Nevertheless, the bookkeeping for the neighbor list and its updating during the optimization make the implementation cumbersome, and therefore a highly simplified expression is used in ROCS implementation[4] and also in several recent shape-feature comparison methods.[8−10] The simplification is to truncate eq 11 to only keep the first summation as

$$V_{AB}^{g} = \sum_{i\in A, j\in B} v_{ij}^{g} \quad (13)$$

This is essentially to truncate the shape density of eq 7 to the first order Gaussian terms: i.e., the shape density of a molecule is expressed as a simple add-up of the shape densities of all atoms in the molecule:

$$G(r) = \sum_{i} g_i(r) \quad (14)$$

This approximation is also called zero-order Gaussian in ROCS,[4] but we think the First Order Gaussian Approximation is a better term because this term is consistent with Grant−Pickup's original Gaussian method. A similar truncation in the hard-sphere model is also used in Phase-Shape.[9] Even though this type of approximation is widely used for its simplicity, the validation of this approach has not been thoroughly studied according to our best knowledge. In a hard sphere model, a molecule's self-overlap volume is equal to the volume of itself. However, the self-overlap volume calculated by First Order Gaussian Approximation (or via the similar hard-sphere approximation) is significantly overestimated in comparison with the hard-sphere model's. Eighteen molecules with diverse chemical structures and molecular size (shown in Table 1) are calculated for self-overlap volumes, and the results are given in Table 1. As shown in the table, the range of the overestimation ratio runs from 1.0 (exact for a single atom molecule, the noble gas Ar) to 6.56 (admantane, which has very compact packing of atom spheres). The large overestimation ratio range not only introduces significant noise in shape similarity calculation but also can lead to incorrect shape alignment due to the unbalanced contribution to shape similarities from different types of functional groups. To illustrate the problem, we align two molecules A and B in two different modes, as shown in Figure 1. The correct shape alignment is Alignment 1 in Figure 1. The FOGA method predicts the best shape alignment is Alignment 2 as shown in Figure 1. The shape Tanimoto similarities for two alignments achieved through different methods are given in Table 2. The reason that FOGA gives incorrect shape alignment is that the alignment of the cyclohexane rings in the two molecules are overemphasized. As shown in Table 1, FOGA overestimates the self-overlap volume of cyclohexane by a factor of 5.7, and this explains why the overlapping of cyclohexane groups are overemphasized in Alignment 2 of Figure 1.

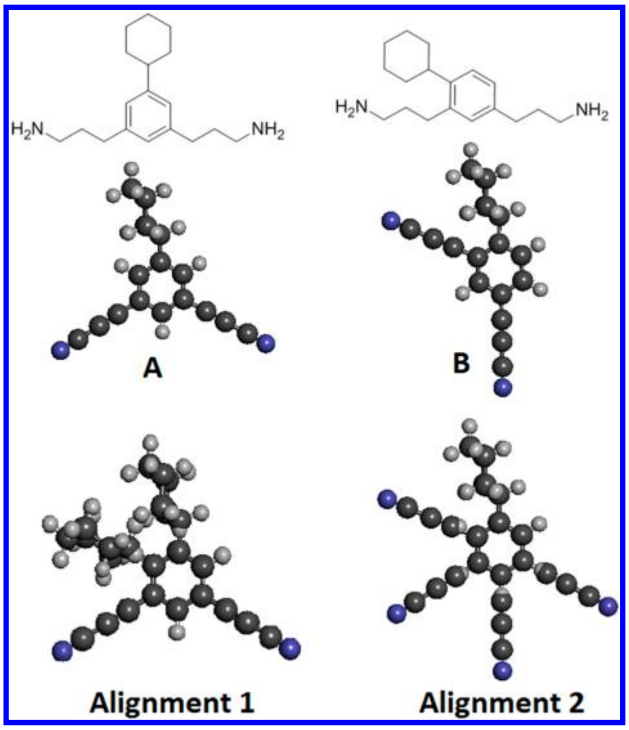


Figure 1. Two different alignment modes of two molecules.

Table 2. Shape Similarity by Different Methods

method	alignment 1	alignment 2
Hard Sphere Model	0.529	0.479
Weighted Gaussian Algorithm	0.530	0.508
First Order Gaussian Approximation	0.478	0.636
Approximate Hard-Sphere Model	0.470	0.627

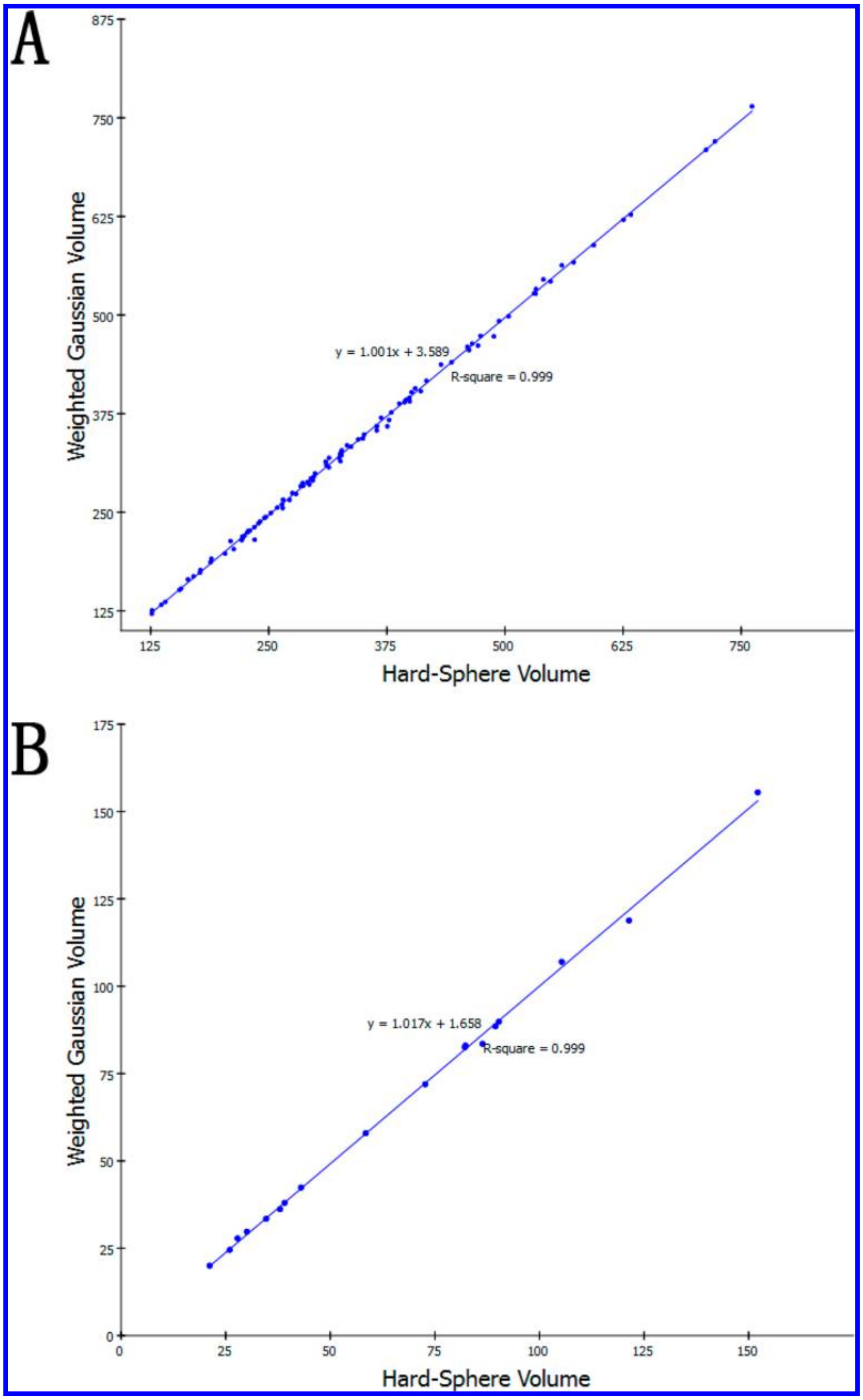


Figure 2. Molecular volumes by Weighted Gaussian Algorithm vs by Hard-Sphere model. (A) For 100 diverse molecules. The universal parameter *k* (0.8665) used in the Weighted Gaussian Algorithm is obtained by fitting Hard-Sphere volumes of 100 diverse ligands with X-ray structures retrieved from PDB. (B) For 18 diverse test molecules from Table 1. These molecules are not used for fitting the parameter *k* and are very different from the CAESAR data set, ranging from the single-atom molecule (Ar) to the highly connected molecule (admantane).

Weighted Gaussian Algorithm (WEGA). A major motivation of this work is to find an effective correction for the overestimation of overlap volumes by First Order Gaussian Approximation while still keeping the efficiency. Since both overlap volumes and self-overlap volumes are used in similarity calculations, one can expect that the accuracy of the overlap

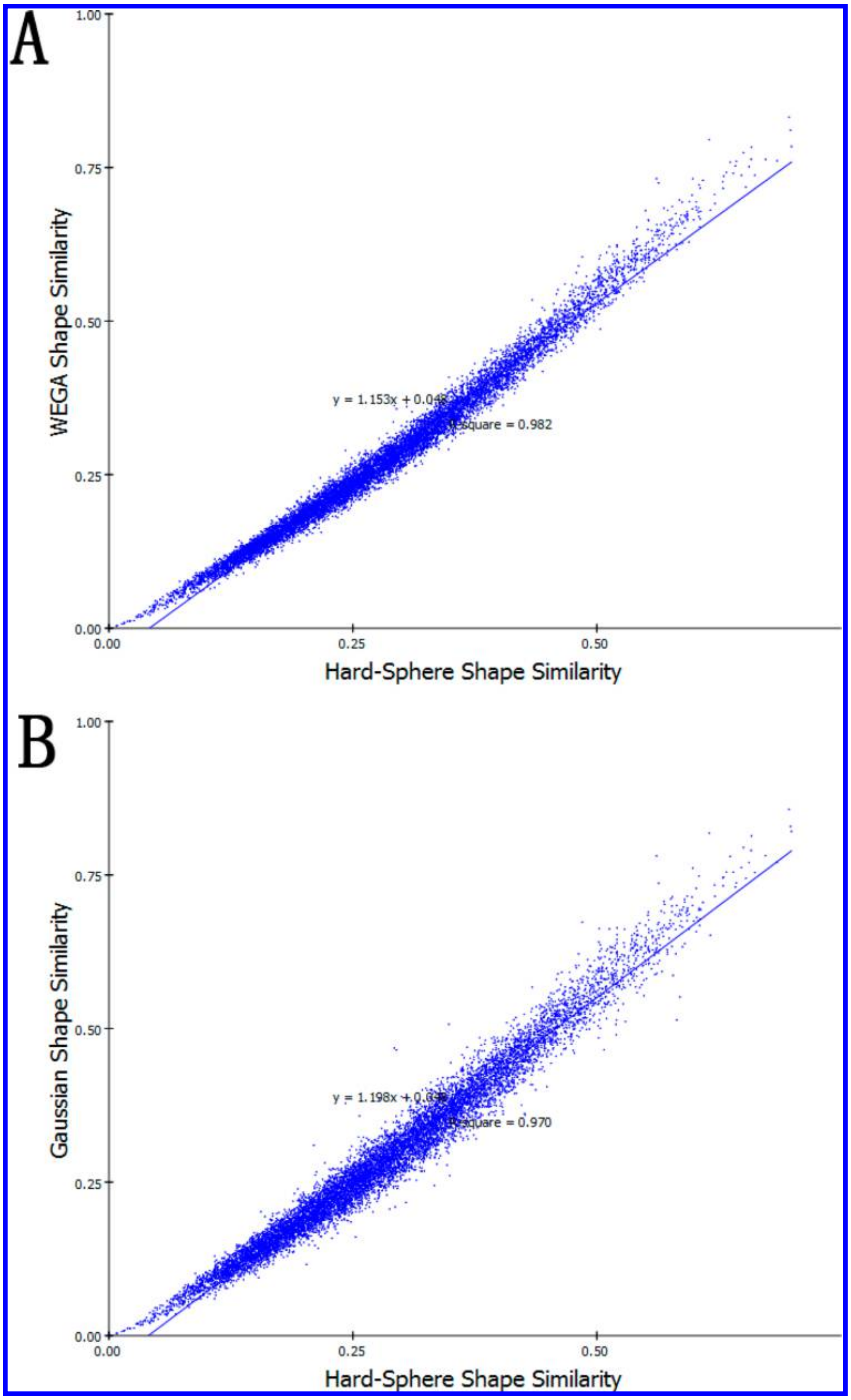


Figure 3. (A) Comparison of shape Tanimoto similarities computed using Weighted Gaussian Algorithm (WEGA) and using Hard-Sphere model. The data show slight nonlinearity. (B) Comparison of shape Tanimoto similarities computed using FOGA and Hard-Sphere model. The data are much scattered around the trend line as compared with that from WEGA in A. The data also show slight nonlinearity.

volumes also has an impact on the accuracy of shape similarity. However, as shown in Table 1, the overestimation ratio has a very large dynamic range ∼1−6.5, therefore a single simple scaling factor does not improve the accuracy of shape similarity calculation. One interesting observation in Table 1 is that the higher ratio molecules have more crowded atoms, which means that the atom overlapping is more significant. For example, in adamantane, the carbon atoms are heavily overlapping their

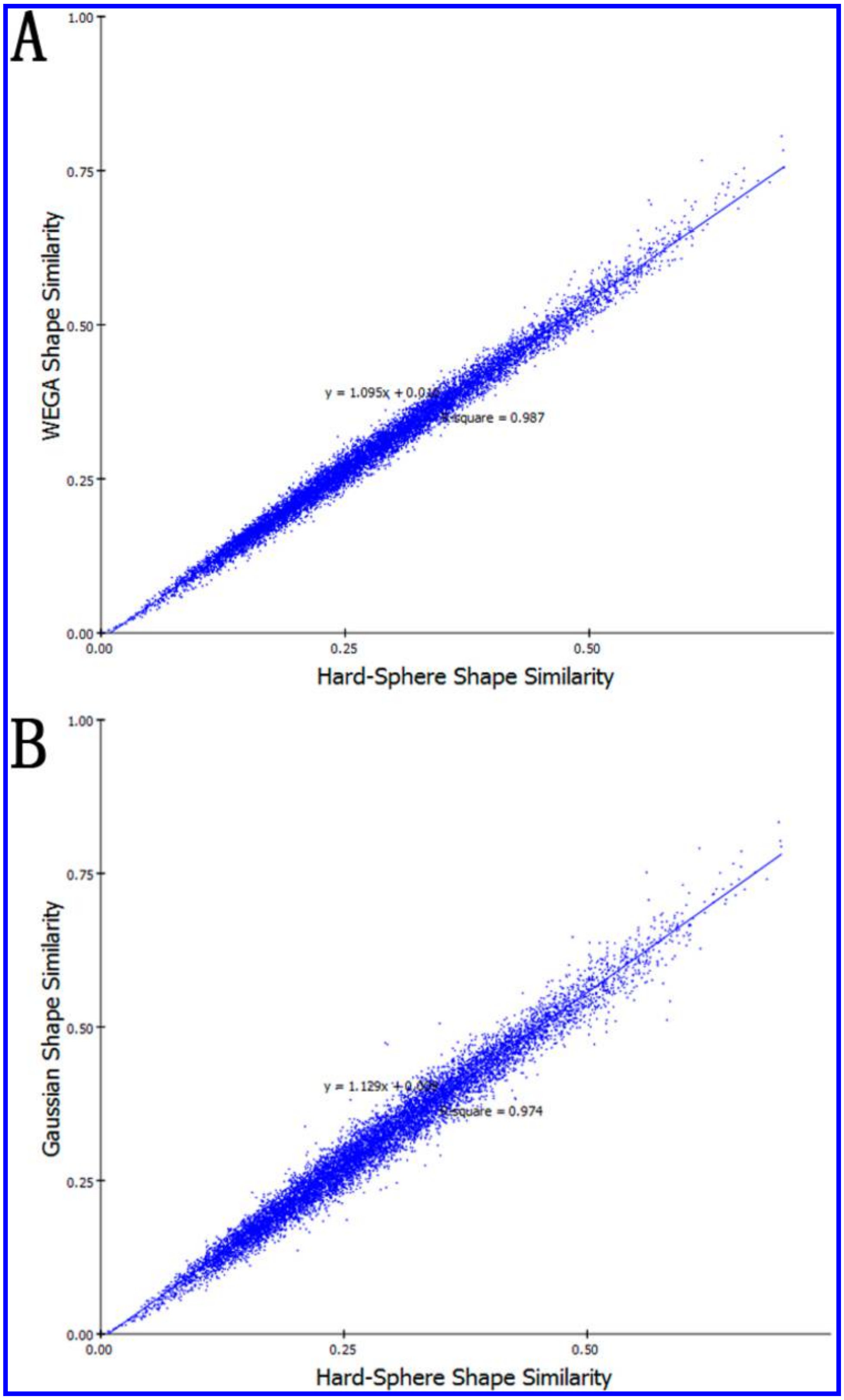


Figure 4. (A) Comparison of shape Tanimoto similarities computed using WEGA and Hard-Sphere model. The linearity is restored by a simple correction which does not change shape similarity ranking in virtual screening. WEGA shows an excellent correlation (r^2 = 0.987) with Hard-Sphere model. The root of mean square (RMS) difference between the WEGA shape similarity and the Hard-Sphere shape similarity is 0.024 for 9900 data points. (B) Comparison of shape Tanimoto similarities computed using First Order Gaussian Approximation (FOGA) and using Hard-Sphere model. The RMS difference of shape similarities computed by FOGA and Hard-Sphere model is 0.038, which is about 60% larger than that of WEGA.

neighbors, and thus this molecule has the highest overestimation for self-overlapping volume, while for the noble gas argon, a single atom molecule, the self-overlapping volume is exactly equal to the hard-sphere volume. Therefore an intuitive correction for the overestimation is to introduce a weight factor for each atom, and the factor is chosen in such a way as to reflect how crowded an atom is with its neighbors in a molecule. For this purpose, we have proposed the Weighted Gaussian Algorithm (WEGA) for molecular shape similarity computation. In this method, the shape density of a molecule is expressed as a linear combination of weighted atomic Gaussian functions as

$$G(r) = \sum_i w_i g_i(r) = \sum_i w_i p e^{-\left(\frac{3p\pi^{1/2}}{4\sigma_i^3}\right)^{2/3}(r-r_i)^2} \tag{15}$$

where w_i is a weighting factor which can be determined by a simple formula. A good measurement for the crowdedness is the total overlap of an atom with all others. Therefore we introduce a simple empirical formula for calculating atomic Gaussian weights:

$$w_i = \frac{v_i^g}{v_i^g + k \sum_{j \neq i} v_{ij}^g} \tag{16}$$

where k is a universal constant determined by fitting the self-overlapping volumes to the hard-sphere volumes for a set of diverse molecules. In this new method, the overlap volume of two molecules becomes

$$V_{\mathrm{AB}}^g = \sum_{i \in \mathrm{A}, j \in \mathrm{B}} w_i w_j v_{ij}^g \tag{17}$$

When the two molecules are identical, the above equation becomes the expression for the self-overlap volume of the molecule, and we want to make the value computed by this equation match the molecule's hard-sphere volume.

RESULTS AND DISCUSSION

Improvement of Molecular Volumes and Shape Similarity. In this study, the ligands from the original CAESAR data set[27] were used for molecular volume calculation assessment, and a subset of 100 diverse ligands from CAESAR data set was used for shape similarity assessment. The subset of diverse ligands was selected by the Find Diverse Molecules protocol in Discovery Studio using all default parameters. The original CAESAR data set contains 913 ligands extracted from the Protein Data Bank, and the structures have been visually checked and some have been corrected by looking into the original papers.[27] In this study, the X-ray conformations of the 100 selected diverse ligands were retrieved from the PDB, and hydrogen atoms were added to the ligands by Discovery Studio. Bondi radii[28] and their recent extension by Truhlar et al.[29] were used for all atoms. The value of k in the above equation is determined by fitting the calculated self-overlapping volumes with the molecular volumes obtained from the hard-sphere model. The best result is obtained for $k = 0.8665$. With the Weighted Gaussian correction, the self-overlapping volumes have nearly perfect correlation ($r^2 = 0.999$) with the hard-sphere volume as shown in Figure 2A. We have also explored some other form of the formula for weight calculations; some of these more sophisticated forms only give negligible improvements for volume estimation. This is understandable since the simple expression for weights works so well, so there is little room for additional improvement. We also tested the volume calculations for a set of 18 molecules shown in Table 1. This set of molecules are not included in fitting parameter k and are very diverse and different from the CAESAR data set. As shown in Figure 2B, the molecular volumes of these molecules by WEGA also accurately reproduce the hard-sphere volumes with the same high correlation ($r^2 = 0.999$). This indicates that the optimal value of k is indeed universal and stable across different data sets.

In order to compare the different methods of computing shape similarity, the same subset of 100 diverse X-ray structures from the CAESAR data set was used. All molecules were translated to the same location with their shape centroids on the same coordinate origin. There are 4950 unique molecular pairs. The overlap volume for each molecular pair is calculated and used to compute the shape similarity. Although the alignment for each pair of molecules is not optimized, the quantities obtained in this way are not true shape similarity, but it is good enough for the purpose of comparing different methods. Similar validations were also used for the Phase-Shape method.[9] To ensure that the validation covers good range of similarity, we also align molecules according to their principal axes of shape so that the largest axis of one molecule is aligned with the largest axis of another molecule for each pair (the same treatment was given to the median and shortest axes). The shape similarity for each pair was then recalculated. This gives an additional 4950 data points. The shape Tanimoto similarity obtained with WEGA is plotted against the results obtained from the hard-sphere model in Figure 3A. For comparison, the similarity obtained from FOGA is also plotted in Figure 3B. Figure 3A shows that the similarity from WEGA has excellent correlation with the hard-sphere results ($r^2 = 0.982$). As indicated by Figure 3B, the similarity obtained from FOGA has larger noise for shape similarity ($r^2 = 0.970$). The data points from WEGA in Figure 3A have a narrow band along the trend line, while the points from FOGA in Figure 3A show a much broader band along the trend line, and some points are scattered far away from the line. Figure 3 also shows some degree of nonlinearity for both methods, and the curvature is quite unique, i.e., the curve bends down in the range 0.0−0.5, and bends opposite in the range of 0.5−1.0 if we use the perfect line of $X = Y$ as the reference. Such a curvature can be easily correctly using an empirical formula $S' = S + 0.03 \sin(2\pi S)$, where S is the original Gaussian shape similarity and S' is the corrected value. One should note that such a modification does not change the ranking by shape similarity scores; thus, it has no impact on the performance for virtual screening. As shown in Figure 4, the curvature is eliminated, and better correlation is shown for both Gaussian methods ($r^2 = 0.987$ for WEGA, and $r^2 = 0.974$ for FOGA). The root of mean square (RMS) errors for the two methods are 0.024 and 0.038 for WEGA and FOGA, respectively. The error of FOGA is about 58% larger than that of WEGA. Figures 3 and 4 are strong indications that the Weighted Gaussian Algorithm not only improves significantly self-overlapping volumes of molecules but also improves the accuracy of shape similarity.

Performance of Volume Calculations. The Weighted Gaussian Algorithm keeps the simplicity of the First Order Gaussian Approximation and can be used to calculate molecular volumes with high accuracy and efficiency. To test the performance for volume calculations, two data sets are used. One is WDI (World Drug Index) version 2010, and the other is the MiniMaybridge database distributed in Discovery Studio software package. The empty records and salts in the WDI data set are removed, and molecules with the number of heavy atoms out of the range 10−64 are also filtered; this gives 70555 molecules from WDI, and one conformation is generated for each molecule in this data set using the CAESAR algorithm[27] in the Discovery Studio.[21] The MiniMaybridge database in Discovery Studio contains 2000 molecules, and 31828 3D conformations are generated using CAESAR. All atoms, including hydrogens, are included in volume calculations. All calculations are done on a 64 bit Linux machine (Intel Xeon CPU, 2.67 GHz). The CPU time and throughput for volume calculations for the two data sets are given in Table 3. About 4000 volumes can be calculated per second for the WDI molecules, with an average of 54 atoms per molecule. For the MiniMaybridge data set, more than 9000 molecular volumes can be computed. The speed of the MiniMaybridge data set is much faster because the molecules have fewer atoms on average. According to eq 17 for volume calculation, the computational cost is proportional to the square of the number of atoms in a molecule. The ratio of speeds for the two data sets is very close to the square of the inverse ratio of their average numbers of atoms per molecules, which is in very good consistency with eq 17.

Table 3. Performance for Molecular Volume Calculation

data set	number of compounds	average number of atoms	CPU time (s)	throughput (volume/s)
WDI	70555	54	17.93	3934
MiniMaybridge	31828	36	3.512	9063

Performance of Shape Similarity Calculations. The shape similarity of a molecular pair is related to maximum overlapping volume of two molecules. To find the maximum overlay of two molecules, the optimization for the molecular alignment is needed. This is usually done via various optimization algorithms, such as the BFGS method used in ref 18. In WEGA, the analytical first and second derivatives of the overlap volume between two molecules with respect to the Cartesian coordinates of atoms can be efficiently calculated and then transformed onto the variables for rigid rotation and translation. Since the computational cost for the second derivatives in the WEGA is only about twice as much as the first derivatives, the Newton−Raphson method is adopted in the optimization of molecule alignment. One practical issue in the optimization of molecular alignment is that multiple local maxima exist for most molecular pairs. Previous investigations suggest that the best shape match can be often found by starting the optimization from the standard orientations.[2−4] To put a molecule into its standard orientation, the eigenvectors of a 3 × 3 shape moment matrix are calculated, and then the molecule is reoriented in such a way so that its largest principal axis is in parallel with the x-axis, the median axis is in parallel with the y-axis, and the shortest axis is in parallel with the z-axis. In WEGA, the shape moment matrix can be calculated according to the following equation:

$$M_{kl} = \sum_{i,j} w_i w_j v_{ij}^g R_{ij}^k R_{ij}^l \tag{18}$$

where R_{ij} is the coalescence center of the Gaussian product of atoms i and j according to eq 20 of ref 2, and the indexes k and l represent the components x, y, or z. Before calculating the matrix, the shape centroid of the molecule is moved to the origin. One should note that the above equation is an approximate expression for computing the principal axes of a molecular shape, and our tests show that such an approximation is adequate for the purpose of initial alignments.

To find the maximum volume overlap of two molecules, we put both molecules into their standard orientation and their shape centroids are superimposed. For each pair of molecules, there are four possible unique initial alignments with their principal axes being coincident. We found that starting from the initial orientations, the optimization of alignment usually converges in less than 10 iterations. The average number of iterations is 7. To test the performance for shape similarity calculations, the same WDI and MiniMaybridge data sets are used to create the conformation databases, and seven query molecules with different numbers of heavy atoms are used to compute the shape similarity with all the database conformations. Similar to ROCS, only heavy atoms are considered to represent the molecular shape, and all heavy atoms use the same van der Waals radii as carbon atom. Both single initial alignment and four initial alignment modes are considered. The CPU timing results for shape similarity calculations are given in Table 4. The speed for shape similarity calculations (in terms of conformations per second) varies with the size of query molecules and the average size of database molecules. For an average size drug-like query molecule (20−25 heavy atoms), the search speed is about 1000−1500 conformations per second in the four-initial alignment mode. For the single-initial alignment mode, the throughput is approximately four times as fast as the four-initial alignment mode.

Table 4. Performance for Molecular Overlay and Shape Similarity Calculations[a]

data set	atoms in query molecules	initial alignments	CPU time (s)	throughput (similarity/second)
WDI	10	4	36.58	1929
	15	4	51.91	1359
	20	4	63.12	1118
	25	4	82.47	856
	30	4	97.98	720
	35	4	107.92	654
	40	4	144.88	486
	10	1	8.80	8014
	20	1	15.67	4503
	30	1	24.49	2882
	40	1	36.08	1956
MiniMaybridge	10	4	11.90	2673
	15	4	16.87	1887
	20	4	20.06	1587
	25	4	26.89	1184
	30	4	32.96	966
	35	4	36.17	880
	40	4	48.92	651
	20	1	5.05	6299
	40	1	12.36	2575

[a] Only heavy atoms are considered in the calculations.

Case Study 1. Overlay of CDK2 (Cyclin-Dependent Kinase 2) Inhibitors. As a simple case study for WEGA, a molecular overlay test was performed for a set of ten CDK2 inhibitors (PDB codes: 1h1q, 1h1r, 1h1s, 1ogu, 1oi9, 1oiu, 1oiy, 2c6k, 2c6m, and 2g9x). The same set was also used in the Phase-Shape performance test.[9] The X-ray structures of the 10 ligands were retrieved from PDB. This data set is selected because it is particularly easy to visually check the quality of the alignment. In the test, both heavy atoms and hydrogen atoms were included in shape-based alignment. The Bondi radii[28,29] were used for all atoms. To find the global best shape alignment for each pair of molecules, four initial alignments are considered for the superposition optimization, and the best one is selected. By default, the first ligand (1h1q) was selected as the target, and all the remaining nine ligands were aligned to it. As shown in Figure 5A, the nine ligands superimpose well with the target purely based on shape. Similar to the ROCS-color, WEGA for molecular overlay can be easily extended to include feature contribution so that one can put weights on both shape and pharmacophore features into account for the alignment. Figure 5B shows the alignment of the 10 ligands using both shape and pharmacophore features. It is obvious from the comparison of Figure 5A and 5B that the shape-feature alignment gives better superposition for pharmacophore features. It is also found that changing to another target ligand, say 2c6k, does not change the quality of the alignment very much. Compared to some recent shape-feature based alignment methods, such as ShaEP,[8] Phase-Shape,[9] and SHAFTS,[10] a major advantage of using Gaussian methods for shape-feature based alignment is that all contribution terms are treated in a consistent way and the analytical first and second derivatives can be obtained easily so that the very robust Newton−Raphson method can be used for the optimization of the alignment.

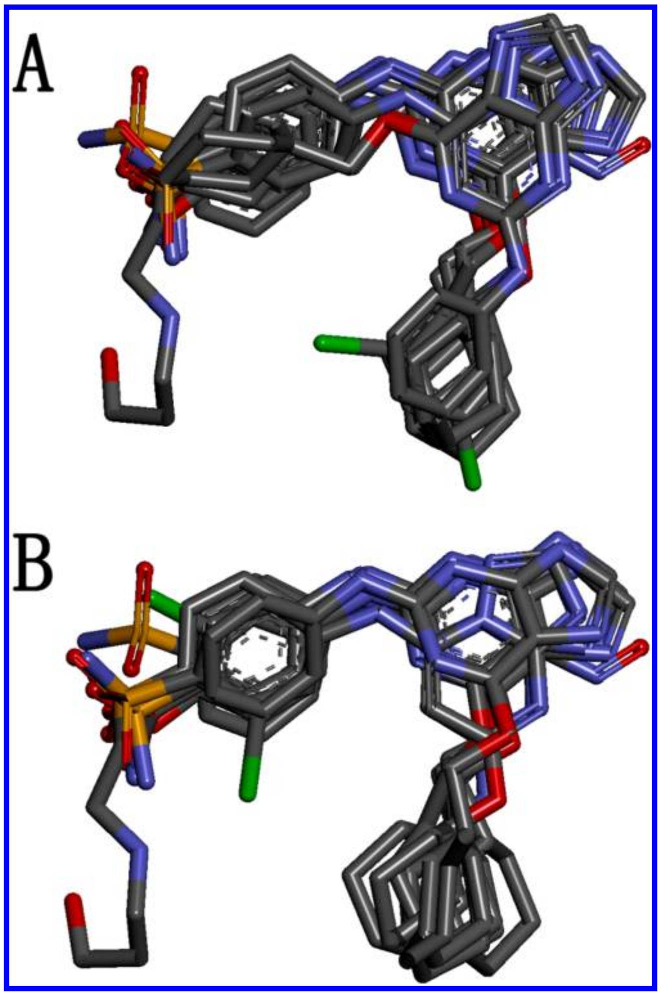


Figure 5. (A) Alignment of 10 CDK2 inhibitors using only shape. (B) Alignment of 10 CDK2 inhibitors using both shape and pharmacophore features. The latter one shows better feature alignment.

Case Study 2: Shape Based Virtual Screening. To test the shape-based virtual screening performance of WEGA, the Directory of Useful Decoys (DUD)[30] database release 2 was used. The original DUD data set has 2950 ligands for 40 different targets. Every ligand has 36 “decoy” molecules that are physically similar but topologically distinct, leading to a database of 98 266 compounds. For each target, the salts and duplicate molecules in its ligand set and “decoy” set were removed by Pipeline Pilot (version 8.5). 3D conformations of ligands and “decoy” molecules were generated using CAESAR algorithm in Discovery Studio (version 3.5). The area under a receiver-operating characteristic (ROC) curve (AUC) was utilized as the performance metric. The query molecules for screening were selected according to the exact procedures by Kirchmair et al.,[31] i.e., the ligands in each target were clustered by maximum dissimilarity using ECFP4 fingerprints[32] in Pipeline Pilot, all ligands in each target were treated as one cluster, and the cluster center was selected as the query molecule. Two scoring functions were used for screening. The first one is pure shape Tanimoto similarity, and the second one is the combination of shape and pharmacophore features (combo scoring). The virtual screening results of the DUD data set based purely on shape similarity computed by FOGA and WEGA methods are shown in Table 5. The shape-feature combo score is the sum of shape similarity and pharmacophore feature similarity. Table 5 also shows the virtual screening results using the combo scoring function computed by FOGA and WEGA methods.

The results of pure shape Tanimoto similarity in Table 5 show that WEGA performs slightly better than FOGA in pure shape based virtual screening. The average AUC for WEGA is 0.728, while FOGA gives an average AUC value of 0.717. Even though the difference is marginal, the general trend for the improvement is clear as 30 out of 40 targets get improved AUC values, and only 9 out 40 targets get slightly worse. One interesting question that we can ask is why the improvement is not significant for pure shape based screening even though the noise of shape similarity score is significantly reduced by using the WEGA method. First, it is clear that pure shape based screening is far from perfect in differentiating the active and the inactive ligands, and some other important factors, such as the interactions between ligands and targets, should be included. Without considering all these factors, the advantage of using the more accurate shape similarity method can be compromised. To explore this idea further, the combo scoring function, which is a combination of both shape similarity score and pharmacophore feature similarity score, was used for screening. Both WEGA and FOGA methods were used to compute combo scores. The screening results are shown in Table 5. As compared with results of pure shape Tanimoto similarity, combo scoring function gives much better results for both WEGA and FOGA method. One interesting point here is that the advantage of using WEGA method vs FOGA method is more obvious when a better scoring function is used for screening. The results of combo scoring function give the same conclusion that WEGA performs better than FOGA; the

Table 5. AUC Values Obtained for 40 Targets in DUD Data Set by Using Both Pure Shape Tanimoto Similarity and Combo Scores of Shape and Features Computed by First Order Gaussian Approximation (FOGA) and Weighted Gaussian Algorithm (WEGA)

target	FOGA (pure shape score)	WEGA (pure shape score)	FOGA (combo score)	WEGA (combo score)
ace	0.302	0.317	0.745	0.819
ache	0.685	0.698	0.783	0.786
ada	0.654	0.716	0.803	0.901
alr2	0.514	0.523	0.523	0.513
ampc	0.903	0.905	0.952	0.962
ar	0.601	0.623	0.593	0.631
cdk2	0.622	0.616	0.819	0.817
comt	0.376	0.394	0.703	0.665
cox1	0.484	0.478	0.533	0.575
cox2	0.965	0.974	0.981	0.981
dhfr	0.772	0.783	0.930	0.959
egfr	0.767	0.747	0.926	0.927
er_agonist	0.765	0.753	0.957	0.962
er_antagonist	0.790	0.795	0.972	0.974
fgfr1	0.572	0.585	0.749	0.828
fxa	0.813	0.826	0.863	0.885
gart	0.916	0.920	0.942	0.938
gpb	0.804	0.843	0.933	0.948
gr	0.734	0.726	0.822	0.870
hivpr	0.669	0.669	0.758	0.801
hivrt	0.486	0.566	0.399	0.413
hmga	0.586	0.597	0.858	0.879
hsp90	0.732	0.751	0.807	0.831
inha	0.736	0.746	0.854	0.864
mr	0.749	0.762	0.818	0.837
na	0.859	0.873	0.978	0.951
p38	0.845	0.835	0.912	0.907
parp	0.673	0.682	0.906	0.948
pde5	0.408	0.484	0.501	0.502
pdgfrb	0.573	0.568	0.641	0.810
pnp	0.867	0.879	0.967	0.931
ppar	0.889	0.898	0.955	0.966
pr	0.651	0.654	0.814	0.864
rxr_alpha	0.916	0.929	0.992	0.994
sahh	0.981	0.969	0.986	0.986
src	0.532	0.549	0.649	0.690
thrombin	0.764	0.786	0.876	0.901
tk	0.676	0.690	0.823	0.826
trypsin	0.815	0.812	0.871	0.897
vegfr2	0.493	0.520	0.554	0.576
mean	0.717	0.728	0.826	0.852

average AUC of WEGA is 0.852, while the average AUC of FOGA is 0.826, which is very close to the average AUC value reported in ref 31 for the 40 DUD targets using ROCS. The gap of the average AUC values between the two methods is 0.026, which is about three times larger than the gap obtained using a pure shape scoring function. The improvement of using WEGA method for combo scoring function is more consistent for the 40 DUD targets. For 24 out of 40 targets, WEGA has better performance with AUC value improved by at least 0.01 comparing FOGA, and only 3 out of 40 targets that WEGA is outperformed by FOGA with AUC values reduced by more than 0.01. This indicates that getting the shape similarity accurate is even more important when we are developing a more accurate scoring function for virtual screening.

■ CONCLUSIONS

The Weighted Gaussian Algorithm (WEGA) provides an efficient and accurate way for calculating molecular volumes, shape similarity, and molecular alignment. The linear combination of weighted atomic Gaussian functions in WEGA keeps the simplicity of FOGA while providing an effective improvement for the accuracy of overlap volumes between molecules. Consequently, the RMS error of shape similarity calculated by WEGA method is reduced to 0.024, while the RMS error of shape similarity by FOGA is 0.038, which is about 60% larger than that of WEGA. About 1000–1500 molecular alignments and shape similarity calculations can be performed for drug-like molecules on a typical one processor machine by this method, thus allowing for large scale virtual screening. Using more accurate shape representation by the WEGA method also improves the accuracy of virtual screening as tested by using all 40 DUD targets. The improvement of virtual screening accuracy is more consistent when the combo scoring function is used. Therefore, WEGA provides a solid foundation for developing more sophisticated and accurate scoring functions for virtual screening. We expect an even wider range of applications for Gaussian function based methods, such as Gaussian based docking,[33] 3D pharmacophore representation,[34] and database searching.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

An additional figure showing the ROC curves of virtual screening of 40 DUD targets using shape-feature combo similarity scores is provided in Figure S1. The First Order Gaussian Approximation (FOGA, in red) and the Weighted Gaussian Algorithm (WEGA, in blue) methods are used for computing the combo similarity scores. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mails: jiaboli@yahoo.com (J.L.), xujun9@mail.sysu.edu.cn (J.X.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Jiabo Li would like thank Drs. Frank Brown and Marvin Waldman for useful discussions. This work was supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2012AA020307)), Guangdong Recruitment Program of Creative Research Groups, the National Natural Science Foundation of China (No. 81001372, 81173470), and the Special Funding Program for the National Supercomputer Center in Guangzhou (2012Y2-00048/2013Y2-00045, 201200000037).

■ REFERENCES

(1) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medical Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862–3886.

(2) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503−3510.

(3) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *14*, 1653−1666.

(4) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. Variable Selection and Model Validation of 2D and 3D Molecular Descriptors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 451−474.

(5) Ballester, F. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.

(6) Morris, R. J; Najmanovich, R. J.; Kahraman, A.; Thornton, J. M. Real Spherical Harmonics Expansion Coefficients as 3D Shape Descriptors for Protein Binding Pocket and Ligand Comparisons. *Bioinformatics* **2005**, *21*, 2347−2355.

(7) Mavridis, L.; Hudson, B. D.; Ritchie, D. W. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787−1796.

(8) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J. Chem. Inf. Model.* **2009**, *49*, 492−502.

(9) Sastry, G. M.; Dixon, S. L.; Sherman, W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlay Scoring. *J. Chem. Inf. Model.* **2011**, *51*, 2455−2466.

(10) Liu, X.; Jiang, H.; Li, H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 2372−2385.

(11) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.

(12) Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116.

(13) Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular Shape Comparison of Angiotensin II Receptor Antagonists. *J. Med. Chem.* **1993**, *36*, 1230−1238.

(14) Lu, W.; Liu, X.; Cao, X.; Xue, M.; Liu, K.; Zhao, Z.; Shen, X.; Jiang, H.; Xu, Y.; Huang, J.; Li, H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 2. Prospective Case Study in the Discovery of Diverse p90 Ribosomal S6 Protein Kinase 2 Inhibitors To Suppress Cell Migration. *J. Med. Chem.* **2011**, *54*, 3564−3574.

(15) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(16) Venhorst, J.; Nunez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of Scaffold Hopping Efficiency by Use of Molecular Interaction Fingerprints. *J. Med. Chem.* **2008**, *51*, 3222−3229.

(17) Rush, T. S., III; J. Andrew Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(18) Haque, I. S.; Pande, V. S. PAPER—Accelerating Parallel Evaluation of ROCS. *J. Comput. Chem.* **2010**, *31*, 117−132.

(19) *FastROCS*, version 1.0; OpenEye Scientific Software Inc.: Santa Fe, NM, USA, 2011.

(20) Putta, S.; Beroza, P. Shapes of Things: Computer Modeling of Molecular Shape in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1514−1524.

(21) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Model.* **1997**, *37*, 80−86.

(22) *Discovery Studio*, version 3.5; Accelrys Software Inc.; San Diego, CA, 2012.

(23) Connolly, M. L. Computation of Molecular Volume. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.

(24) Gibson, K. D.; Scheraga, H. A. Exact Calculation of the Volume and Surface Area of Fused Hard-Sphere Molecules with Unequal Atomic Radii. *Mol. Phys.* **1987**, *62*, 1247−1265.

(25) Richmond, T. J. Solvent Accessible Surface Area and Excluded Volume in Proteins. *J. Mol. Biol.* **1984**, *178*, 63−89.

(26) Boys, S. F. Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proc. R. Soc. London* **1950**, *A200*, 542−554.

(27) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'Brien, S.; Kirchmairs, J. CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923−1932.

(28) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(29) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. Consistent van der Waals Radii for the Whole Main Group. *J. Phys. Chem. A* **2009**, *113*, 5806−5812.

(30) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(31) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. *J. Chem. Inf. Model.* **2009**, *49*, 678−692.

(32) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(33) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76−90.

(34) Taminau, J.; Thijs, G.; Winter, H. D. Pharao: Pharmacophore Alignment and Optimization. *J. Mol. Graphics Modell.* **2008**, *27*, 161−169.