ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential

Mikko J. Vainio,* J. Santeri Puranen, and Mark S. Johnson

Structural Bioinformatics Laboratory, Department of Biochemistry and Pharmacy, Åbo Akademi University, Tykistökatu 6A (BioCity), FI-20520 Turku, Finland

Received September 2, 2008

ShaEP is a tool for rigid-body superimposition and similarity evaluation of ligand-sized molecules. Molecular overlay methods traditionally work on either substructures, molecular surfaces or interaction fields, or atom-centered Gaussian functions representing the molecular volume. While substructure searches are unlikely to reveal hits that are chemically different from the template structure, the other methods are capable of "scaffold hopping". Methods that match characteristic points in interaction fields can find alignments in situations where only some portions of the structures match but potentially miss good alignments if the used point sets are not detailed enough, which in turn increases the runtime of the used graph algorithms beyond practical limits. The faster, polynomially scaling volumetric methods consider the whole space to be equally important, which works well for molecules of equal size but partial matches might go undetected. ShaEP aims to capture the strengths of both field-based and volumetric approaches. It generates initial superimpositions using a matching algorithm on graphs that coarsely represent the electrostatic potential and local shape at points close to the molecular surfaces. The initial alignments are then optimized by maximization of the volume overlap of the molecules, computed using Gaussian functions. ShaEP overlays drug-sized molecules on a subsecond timescale, allowing for the screening of large virtual libraries. The program is available free of charge from www.abo.fi/fak/mnf/bkf/research/johnson/software.php.

1. INTRODUCTION

Identification of new lead molecules by assessing their similarity to an existing active molecule has been a widely studied and used method in the development of pharmaceuticals. Several different methods have been devised for the detection of similarity based on molecular shape and/or electrostatics,[1−15] pharmacophore points,[16] and/or other abstracted mathematical descriptions of the molecular shape.[17−19] Many of these methods are devised to align the compared structures via maximization of the similarity function that is used. For reviews of molecular similarity methods, see refs 20− 22. A table summarizing the different methods is presented in ref 23.

In a recent study, it was shown that drugs that bind to the same target tend to exhibit a high level of 2D similarity.[24] The authors stated that this tendency has no physical basis but is a consequence of inductive bias in the process of human reasoning during the development of drugs. This notion motivates the use and development of virtual screening methods based on molecular similarity in three dimensions, which seek to identify ligands that would bind to the biological target but avoid patented chemistry. In addition, novel molecular scaffolds may lead to drugs with improved pharmacokinetic and safety profiles.

Many of the 3D similarity methods consider the molecular electrostatic potential field (MEP).[3,4,13,14,19,25,26] Calculations on the MEP, expressed as a rectangular or spherical grid around the molecule, may involve thousands or tens of thousands of grid points causing these calculations to be very time-consuming. Consequently, the MEP has been approximated by using, e.g., local extrema of the MEP,[4,25] but finding the extrema still requires the calculation of the full MEP. Recently Cheeseright et al.[14] introduced a method that extracts points of extrema from molecular interaction fields, while the evaluation of the full grid is avoided.

In this study, a method is presented for the detection of similarity in molecules and the subsequent rigid-body superimposition of their structures. The method, called ShaEP (reminiscent of Shape and Electrostatic Potential), uses a maximal common subgraph isomorphism algorithm on graphs whose vertices (nodes) are colored (labeled) with the electrostatic potential (ESP) calculated at points projected from the molecular geometry. Each of the subgraph matches is then used to calculate the geometric transformation that superimposes the matched vertices. Because many of the subgraph isomorphisms differ only slightly (e.g, swap of two vertices close to each other in space), the geometric transformations show a degree of degeneracy. A clustering algorithm is applied to produce a set of nondegenerate transformations, which is then used to superimpose the molecular structures. The superimposition is scored according to the shape-density overlap volume calculated using a Gaussian description of molecular shape.[27,28] Gaussians are frequently applied in molecular superimposition algorithms.[1,7−9,15,29,30] ShaEP bears similarity to the field-graph method described by Thorner et al.[4,25] but differs in the generation of the graph (placing of the vertices), comparison of the labels (ESP and a local shape descriptor) of the vertices, and in the calculation of the final similarity values. Especially, in ShaEP the full grid representation of the MEP

* Corresponding author phone: +358-2-215 4600; e-mail: mikko.vainio@abo.fi.

10.1021/ci800315d CCC: $40.75
Published on Web 02/05/2009

Chart 1. See Chart 1 for algorithm 1.

Algorithm 1 The ShaEP algorithm

Require: 3D model(s) of a template and one or more target molecules
1: Build a field-graph for each template molecule (see section 2.1)
2: **while** not(Termination()) **do**
3: Read next target molecule structure
4: Build a field-graph for the target molecule
5: **for** Each template molecule **do**
6: Find (near) maximal subgraph matches of the field-graphs (section 2.2)
7: Calculate the rigid-body transformation required to superimpose each subgraph match
8: Cluster similar transformations (section 2.3)
9: Optimize the superimposition defined by each of the remaining transformations (section 2.4)
10: Record the highest similarity score for the target molecule
11: **end for**
12: Update hitlist with the target molecule
13: **end while**
14: Output hitlist

is not calculated during the process, which results in a fast construction of the field-graph.

The field-graph method is known to miss approximately 6% of the known similar compounds.[3] There are several reasons for this, such as the use of an insufficient set of field-graph vertices,[3] insufficient conformational sampling, and poor description of the MEP due to the use of an overly coarse partial atomic charge model (used in high-throughput applications). However, the subgraph isomorphism may reveal similar regions in the MEPs in the situation where the overall molecular sizes do not match.

The main application of ShaEP is in the virtual screening of large databases of compounds. For the majority of these compounds, the bioactive conformation is not known, but a conformer ensemble must be used as input to similarity analysis. ShaEP addresses the problem of conformational flexibility via the sequential processing of multiple structures for each database molecule: consecutive structures with the same name are considered as conformations (these may also include different protonation states or tautomers) of the same compound, and the maximum of the computed similarity indices over these alternate representations is reported for the compound. Multiple structures can be used for the template compounds as well, which causes ShaEP to superimpose each conformer of a database compound on each of the template structures and report the maximum similarity index obtained over all of the comparisons. By using a pregenerated conformer database as input to the program, a flexible similarity analysis is achieved without incurring the repeated conformer generation on each run.

A retrospective virtual screening experiment shows that ShaEP is capable of identifying a substantial number of active compounds in a database of druglike molecules.

The program is available free of charge from www.abo.fi/fak/mnf/bkf/research/johnson/software.php.

2. METHODS

2.1. Generating Field-Graphs. The 3D models of the molecules are used to calculate points in space at which the ESP will be evaluated. These are the vertices of the field-graph *G*. In contrast to Thorner et al. who extracted characteristic points in the MEP grid using a clustering algorithm,[25] there is nothing particularly characteristic about the chosen locations, but they merely serve as a deterministically selected set of reference points for detecting similarities in the MEPs of the two molecules. Vertices are added around each atom according to the atom's hybridization state and simple geometric rules: For atoms connected by a single bond to exactly one non-hydrogen neighbor, a vertex is placed in the direction opposite to the non-hydrogen neighbor at a distance of $\sigma + h$ from the origin atom, where σ is the van der Waals radius of the atom and h is a user-adjustable constant that by default takes the value of 0.2 Å. The following rules apply only to non-hydrogen atoms. Add vertices in the place of "missing" neighbor atoms: sp^3-hybridized atoms should have four neighbors in a tetrahedral geometry, sp^2-hybridized atoms should have three neighbors in a planar geometry, and linear sp-hybridized atoms should have two neighbor atoms. Again the vertices are placed at a distance of $\sigma + h$ from the origin atom. Figure 1 illustrates the vertices generated for furan.

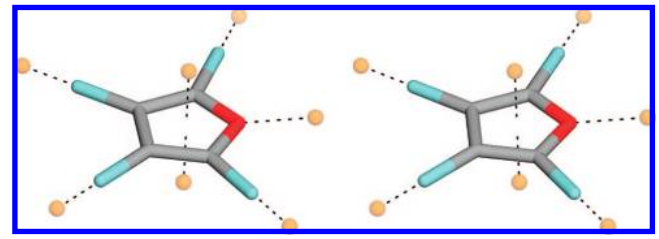

Figure 1. An example of a field-graph generated for furan. The number of field-graph nodes generated for a compound is of the same order as the number of atoms in the molecule, which allows for the perception of graph isomorphism between field-graphs in a reasonable time. See the text for details. The wall-eyed stereo figure was generated using PyMOL.[31]

Additional vertices are added on both sides of (nearly) planar rings with more than four atoms. A principal component projection is used to find the orthogonal directions of variance in the coordinates of the atoms of the ring. The principal axes are the eigenvectors of the covariance matrix of the coordinates. The axis corresponding to the smallest real eigenvalue of the analysis is the normal of the plane of the ring. The quality of the fit is given by $1 - \min(eigenvalues)/\max(eigenvalues)$; a value of one means a perfectly planar ring (no variance in the direction of the normal to the plane). If the quality of the fit is higher than a threshold value (by default 0.98), vertices are added on both sides of the ring at a distance of 1.6 Å along a line that is parallel to the normal of the ring plane and extending through the centroid of the ring.

Once all vertices are generated, those vertices that collide with any atom of the molecule (vertices within the van der Waals radius for the atom) are rejected. The remaining vertices that are closer to each other than a distance threshold (by default 1.0 Å) are clustered together using a form of maximal linkage clustering (as described below for geometric transformations in section 2.3).

Each of the vertices remaining after clustering is assigned two labels: the ESP and a local shape descriptor. The ESP is computed in volts (SI units) using a simple Coulombic function

$$\varphi_E = \frac{1}{4\pi\epsilon_0\epsilon_r}\sum_i \frac{q_i}{d_i} \quad (1)$$

where q_i is the partial charge of atom i in Coulombs, d_i is the Euclidean distance between the atom and the vertex in meters, ε_r is the relative static permittivity of the medium (defaults to one for vacuum; user-adjustable parameter),

$1/(4\pi\epsilon_0)$ is Coulomb's constant, and the summation runs over all atoms in the molecule.

The local shape descriptor is a histogram vector of signed distances of atoms from a plane tangential to the molecular shape-density (surface) at the vertex coordinates $\mathbf{r} \in \mathbb{R}^3$. The shape-density of atom i at $\mathbf{r}$ is expressed as a spherical Gaussian

$$\rho_i(\mathbf{r}) = p_i \exp(-\alpha_i(\mathbf{r} - \mathbf{R}_i)^2) \quad (2)$$

where $\mathbf{R}_i$ is the atomic coordinate. The amplitude is set to $p_i = 2\sqrt{2}$, and the decay factor α_i is calculated from p_i and the van der Waals radius of the atom σ_i as in ref 27

$$\alpha_i = \pi\left(\frac{3p_i}{4\pi\sigma_i^3}\right)^{2/3} \quad (3)$$

The molecular shape-density at the vertex location is the sum of all individual atomic densities. The gradient of the molecular shape-density, obtained from basic algebra, is opposite in direction to the normal of the tangential plane at that location. Given the normal of the plane $\mathbf{n}$, the signed distance of atom i to the plane is $\mathbf{n}\cdot\mathbf{R}_i - \mathbf{n}\cdot\mathbf{r}$. These distances for atoms within a given radius (by default 10.0 Å) from the vertex are included in the histogram vector using a default bin size of 1.0 Å. Negative distances are counted in the bins in the first half of the vector and positive distances in the bins of the second half. After counting the distances, the histogram vector is normalized to a unit Euclidean norm (length).

2.2. Graph Matching. The labeled vertices constitute a completely connected graph, i.e., there is an edge connecting every pair of vertices. Each edge is labeled with the Euclidean distance between the vertices it connects. The backtracking search algorithm of Krissinel et al.[32] is then used to find maximal subgraph isomorphisms between the graphs of the template G_A and the target G_B molecule. The isomorphism algorithm finds and reports subgraph matches that contain at least n_0 (a user-adjustable parameter) nodes and works faster when n_0 increases. During the matching, the value of n_0 is updated after each detected isomorphism to the number of nodes in the largest subgraph match found so far. This allows for a relatively small initial value of n_0 without incurring the cost of finding all the small subgraph matches in cases where much larger (and more informative) matches exist. The subgraph isomorphism matching has a worst-case scaling of $O(m^{n+1}n)$, where $m = \|G_A\|$ and $n = \|G_B\|$ are the number of nodes in the two graphs,[32] and becomes the most time-consuming part of the ShaEP algorithm if the sizes of the graphs grow beyond a few tens of nodes.

During the matching, the labels of the vertices and edges are compared for compatibility. In order for two nodes to match, their ESP may differ at most by 0.5 V (a user-adjustable parameter), and the dot product of their shape descriptor vectors must yield a value ≥ 0.866 (a user-adjustable parameter). The dot product equals the cosine of the angle between the unit vectors; thus, the latter condition translates into an angle less than 30°. In order for two edges to match, the difference in the distance between the connected nodes in the two graphs must be ≤ 1.0 Å (a user-adjustable parameter).

The detected subgraph isomorphisms are filtered in order to remove the least informative matches involving less than 90% of the number of vertices of the largest match. Each of the remaining isomorphisms gives a list of matching points in $\mathbb{R}^3$ that is then used to calculate an optimal geometric transformation (rigid-body motion) of the target vertices that superimposes them on the template vertices in a least-squares sense.[33] (In order to calculate the superimposition in 3D the minimum number of nodes in a match n_0 must be ≥ 3.)

The set of transformations computed from the isomorphisms is augmented with four transformations that superimpose the geometric centroids and align the principal axes of the two structures. One of the transformations aligns the principal axes, and the other three transformations correspond to a 180° flip of the target structure about each of its principal axes in turn.

2.3. Clustering Transformations. The set of geometric transformations that is obtained is redundant in the sense that many of the transformations, when applied to the target structure, result in very similar superimpositions. Therefore, the transformations are clustered and replaced with the average transformation for each cluster using the following maximal linkage style algorithm: Generate a vicinity graph where the transformations are vertices and an edge exists between two vertices if the distance between the transformations is below a tolerance value (the distance measure is defined later). Then, iteratively find the maximal clique (the largest possible subgraph where each vertex has an edge to every other vertex in the subgraph) in the vicinity graph, replace the clique vertices in the graph with their average (defined later), and repeat until the size of the maximal clique is one, i.e., there are no edges left in the vicinity graph. The remaining vertices are the cluster centroids, a set of nonredundant transformations. We use the branch-and-bound algorithm of Tomita and Kameda[34] (an almost identical algorithm was coincidentally described by Konc and Janežič)[35] for clique detection. Analogous clustering schemes have been used previously in molecular similarity calculation programs.[10,36]

The same clustering procedure is used on the set of initially generated field-graph vertices above. The regular Euclidean distance is used for field-graph nodes for which we have coordinates in $\mathbb{R}^3$, but in the case of transformations that are combinations of a rotational and a translational movement, no unambiguous distance metric exists.[37] The usual solution to this problem is to use a weighted sum of some distance measure for the rotational and translational parts of the movement (specific for the clustering of transformations, see ref 10). While a weighted sum works well for detecting similar transformations, computing the average of a set of transformations is less straightforward. Fortunately, this problem has received attention in robotics and computer graphics research, and recently Kavan et al. formulated a method for averaging transformations using dual quaternions.[38]

Unit quaternions are routinely used in computer graphics and molecular modeling to represent rotations about an axis that goes through the origin of the coordinate system (see ref 39 for a review on the use of quaternions in molecular modeling). A unit *dual* quaternion represents a screw motion, that is a simultaneous rotation and translation along an arbitrary axis in $\mathbb{R}^3$. Every rigid transformation can be represented as a screw motion.[37] Dual quaternions $\hat{\boldsymbol{q}}$, first described by Clifford in 1882,[40] are dual numbers of ordinary

quaternions $\mathbf{q}$: $\hat{\mathbf{q}} = \mathbf{q}_0 + \epsilon\mathbf{q}_\epsilon$, where the dual unit ϵ satisfies $\epsilon^2 = 0$. A treatment of the algebra of dual quaternions, much of which can be implemented using the algebraic operators of the component quaternions, is beyond the scope of this paper; the introduction given by Kavan et al.[38] suffices in order to repeat our work. Worth mentioning is the antipodal property of unit dual quaternions, that is, $\hat{\boldsymbol{q}}$ and $-\hat{\boldsymbol{q}}$ represent the same rigid transformation (rotating an angle ϕ and translating an amount t about an axis $\mathbf{a}$ yields the same result as rotating by $-\phi$ and translating by $-t$ about $-\mathbf{a}$). Obviously, the antipodality must be accounted for by any distance measure used between dual quaternions. The ordinary quaternions are also antipodal; therefore, the distance measures defined for the ordinary quaternions can be used to address the antipodal equivalence problem of dual quaternions as well. We use the angular distance measure, formulated originally for quaternions by Park et al.,[41] as the distance between dual quaternions

$$d = \min(|||\log(\hat{\mathbf{q}}_1^{-1}\hat{\mathbf{q}}_2)||, |||\log(\hat{\mathbf{q}}_1^{-1}(-\hat{\mathbf{q}}_2))||) \quad (4)$$

where $\hat{\mathbf{q}}_1{}^{-1}$ denotes the inverse of a dual quaternion. The default distance tolerance value is 2.0. The average of the transformations in a clique is calculated using the dual quaternion iterative blending algorithm of Kavan et al.[38] With the distance and average operations defined, the clustering of transformations is accomplished.

2.4. Overlay Optimization. The transformations remaining after the clustering are applied to the target molecular structure. The superimposition is scored according to the overlap of the shape-densities of the template (A) and target (B) structures and their field-graphs. In this study, the shape-density overlap V of two molecules is defined as the sum of the overlap integrals of individual atomic shape-densities[9]

$$V = \sum_{i\in A}\sum_{j\in B}\int d\mathbf{r}\rho_i(\mathbf{r})\rho_j(\mathbf{r}) \quad (5)$$

$$= \sum_{i\in A}\sum_{j\in B} p_i p_j \exp\left(-\frac{\alpha_i\alpha_j d_{ij}^2}{\alpha_i+\alpha_j}\right)\left(\frac{\pi}{\alpha_i+\alpha_j}\right)^{3/2} \quad (6)$$

where i and j run over the template and target structures, the integral is over the whole space, and d_{ij} is the Euclidean distance between the atomic centers $\mathbf{R}_i$ and $\mathbf{R}_j$. For further reference on the use of Gaussians to represent the molecular shape, see refs 9, 27, 28, and 30.

The amplitudes p of the Gaussians in eq 6 can be weighted according to some physicochemical property of the atom, such as hydrophobicity, hydrogen-bonding potential, electrostatic potential, or partial charge, and the weighted volume overlap can thus be used to obtain a pharmacophore-style alignment.[1,7,9,42] However, the atom-centered properties might not convey as biologically relevant information as the field around the molecule. Inspired by Cheeseright et al., who used molecular field extrema as a basis for superimposition,[14] ShaEP uses a set of Gaussians placed at the field-graph vertices to evaluate a "volume" of the field-graph overlap. The overlap integral of these Gaussians is weighted according to the difference in the ESP at the corresponding vertices according to

$$V_{E,AB} = \sum_{k\in G_A}\sum_{l\in G_B}\exp(-\beta||\varphi_{E,k} - \varphi_{E,l}||)\int d\mathbf{r}\rho_k(\mathbf{r})\rho_l(\mathbf{r}) \quad (7)$$

where k and l run over the field-graph vertices of the template and target structures; a radius of 2.0 Å is used to obtain the decay factor α for the Gaussians, $\beta = 1$ by default, and otherwise the coefficients are as for eq 2. The weighting term is never negative, thus V_E does not penalize for the overlap of field-graph vertices whose ESP is of opposite sign but only rewards for the overlap of similar regions.

The overlap volume can be differentiated with respect to the rotation and translation components of the rigid transformation required for superimposition.[1,43] The analytical first and second derivatives allow for rapid optimization of the overlap volume. We adapted the TNPACK truncated Newton optimization routine of Schlick and co-workers.[44−47] The objective function is the sum of eqs 6 and 7. For efficiency, the Gaussians corresponding to apolar hydrogens (taken as those bonded to carbon atoms without reference to any computed atomic property) are omitted from the summations in eq 6 during the optimization. The full overlap integral, including Gaussians for all hydrogens, is evaluated only once after the optimization has converged.

Each geometric transformation in turn is used as a starting point for the maximization of the sum of eqs 6 and 7. The superimposition that gives the largest objective function value is used to calculate the final similarity index. Another option, which is not pursued in this study, would be to compute the average similarity over a simultaneous superimposition on multiple prealigned template molecules. This approach has been shown to improve selectivity over using a single template molecule.[48]

The value of eq 7 depends on both the difference in ESP and the distance between the field-graph vertices. While eq 7 is economical to use during optimization (no need to recalculate the ESP), the final similarity score must be independent of the distance between field-graph vertices because we want to measure the similarity of the fields, not the graphs—the locations of the vertices have origins in the molecular geometry for which the shape overlap already accounts for. Therefore, the ESP of molecule B φ_E^B is evaluated at the location of each field-graph vertex of molecule A (vertices within the van der Waals radius of any atom of molecule B are omitted) and vice versa

$$V'_{E,AB} = \sum_{k\in G_A}\exp(-\beta||\varphi_{E,k} - \varphi_E^B(\mathbf{r}_k)||) + \sum_{l\in G_B}\exp(-\beta||\varphi_{E,l} - \varphi_E^A(\mathbf{r}_l)||) \quad (8)$$

ShaEP uses the weighted average of the normalized overlap volume computed using eq 6 and the normalized value of eq 8 as the final score value. The value of eq 6 can be normalized by using the overlap integrals of the molecules A and B with themselves.[28,30] The final score is expressed as the Hodgkin[49] similarity index

$$S_{AB} = \frac{wV'_{E,AB}}{||G_A|| + ||G_B||} + \frac{(1-w)2V_{AB}}{V_{AA}+V_{BB}} \quad (9)$$

where w is a relative weighting factor, by default $w = 0.5$.

2.5. Test Runs. *2.5.1. Reproduction of Superimpositions of X-ray Structures.* There are several different potential performance metrics for a molecular similarity/alignment algorithm. Here, we consider the ability to reproduce the correct alignment of ligands as obtained by superimposing multiple X-ray crystal structures of a protein that contain different cocrystallized ligands, extracted from the Protein Data Bank (PDB),[50] and the ability to distinguish between active and inactive molecules in a database screen.

A molecular superimposition/similarity analysis tool should be able to reproduce alignments of ligands found in X-ray crystal structures of protein−ligand complexes—if the algorithm does not consider the same aspects of the ligands important as the host biomolecule does, there is little hope that the algorithm will succeed in virtual screening either. Moffat et al. called this type of test a "sanity check".[51] In the usual case, structures of protein−ligand complexes are superimposed based on the Cα positions of the protein. The orientation of a ligand in the superimposed X-ray structure is then used as a reference for the computation of heavy-atom rmsd of orientations of the ligand in superpositions generated by the algorithm. The rmsd is a well-understood measure among the community, but care must be taken not to draw overly "accurate" conclusions on the obtained rmsd values given the experimental accuracy of the crystal structures.

The diffraction-component precision index (DPI), introduced by Cruickshank[52] and later simplified by Blow,[53] has been used as a measure for the experimental accuracy of the atomic coordinates in an X-ray crystal structure,[54,55] and its wider use in the evaluation of virtual screening methods has recently been promoted by Hawkins et al.[56] and Nicholls.[57] In this study the DPI is computed according to the formulation by Goto et al.[54]

$$\sigma(r, B_{\mathrm{avg}}) = 2.2 N_{\mathrm{atoms}}^{1/2} V_{\mathrm{a}}^{1/3} n_{\mathrm{obs}}^{-5/6} R_{\mathrm{free}} \qquad (10)$$

where N_{atoms} is the number of atoms with full occupancy in the asymmetric unit, V_{a} is the volume of the asymmetric unit (unit cell volume divided by the number of asymmetric units), and n_{obs} is the number of unique reflections used in the refinement. The program used to calculate the DPI values along with its C++ source code is available under an open-source license from the ShaEP Web site. Equation 10 gives the approximate standard error in atomic positions, thus for the rmsd the error estimate is $\sqrt{2}\sigma$. Goto et al. considered an rmsd $> 2\sqrt{2}\sigma$ as a significant deviation.[54] We consider a computed superposition that has an rmsd greater than 2.0 Å with its X-ray structure to be "different". For nearly all the studied structures $\sqrt{2}\sigma$ is lower than the selected limit.

The structure data sets of ligands, where the ligands within each data set are cocrystallized with the same protein, are listed in the Supporting Information. No diversity analysis was performed at this stage of selecting structures, only duplicate ligand molecules were avoided. The structure with the lowest DPI value was preferred when selecting between multiple models of the same molecule. In all cases, the crystal structures were aligned on the chain indicated in the Supporting Information using the protein structure alignment program VERTAA[58] (available in the program Bodil[59]). Ligands were extracted from the superimposed crystal structures, bond orders and formal charges (deprotonated carboxylate groups, protonated primary amines) were manually assigned, and hydrogen atoms were added to the models using Balloon.[60] The positions of the added hydrogen atoms were not optimized by any means. The manual corrections were made using Maestro.[61] MMFF94-like partial atomic charges were assigned to the ligands using Balloon.

The selected ligands within each set do not all form the same contacts with the host protein or occupy the same regions of the binding pocket. Furthermore, some pairs of compounds in these sets are trivially easy to superimpose because the structures are closely related to each other, and some pairs are very difficult because the structures have little in common. Therefore, pairs of ligands must be selected for benchmarking using a procedure that discards very similar ligand structures from consideration, ensures the "sanity" of the check, i.e., that there exists a meaningful correct solution for each pair of superimposed compounds, and is independent of the superimposition method under scrutiny. We used the program Contactos[62] to produce an all-against-all matrix of similarity values for the ligands within each set (using option '-t' in Contactos). Contactos computes a similarity score between two ligands based on the contacts made between ligand atoms and atoms of the host protein. The similarity score has a minimum of zero (no mutual contacts) and a maximum of one (matching contacts). The same reference protein structures were used as above for the superpositioning of the X-ray structures using VERTAA. ShaEP was then run using each ligand as the template structure in turn, while the target structures were selected to have a similarity score within the range [0.3, 0.95] with the template structure as computed using Contactos. The use of these similarity cutoffs discards unfeasible cases from consideration. As a general remark, some of the ligand sets, e.g. HIV protease ligands, are not suitable for benchmarking a program that performs conformation sampling of the structures because of the overwhelming conformational freedom of the ligands (see also ref 63 for criticism of the use of HIV protease ligands as a benchmark data set). Because ShaEP operates on rigid structures, these conformationally very flexible ligands were included in the test runs.

For each ligand orientation produced by ShaEP, the rmsd of non-hydrogen atoms was calculated against the X-ray reference orientation for the ligand using in-house software that takes into account the possible topological (2D) symmetry of the ligand.

2.5.2. Virtual Screening. ShaEP was used to perform a virtual screening on the data sets of Jain (titled "*JMedChem 2004 Test Data (rev1)*", downloaded from www.jainlab.org).[64] The data consist of sets of active compounds for four different biological targets and two decoy sets of inactive druglike compounds. Of these two decoy sets, we used the one collected from the ZINC database,[65] which has been regarded as a more challenging "background noise" against which the active molecules should be detected in virtual screening studies.[66] Balloon was used to generate conformer ensembles for the ligands and the decoys. The settings for Balloon were as follows: population size of 20 conformers and 200 allowed generations for the genetic algorithm and 20 iterations of conjugate gradient geometry optimization for each conformer in the final conformer ensemble. The other settings were kept at their default values. The procedure assigns MMFF94-like partial atomic charges to the generated conformers. ShaEP was then run on the data sets using both the full algorithm and using only volume overlap optimization without the field-graph steps (option *onlyshape*). The ligand structures from the "best automatically generated hypothesis" were used as template structures for each biological target (file "ACSHypo.mol2" in the corresponding directory).

It can be trivially easy to separate active molecules from inactive ones in a virtual screening benchmark data set if care is not taken to include inactive decoy compounds that

are of close chemotype to the active structures or if the active structures are members of an analogous series of compounds.[24] The DUD collection of actives and decoys,[67] the largest freely accessible data set for benchmarking virtual screening methods published to date, was used to test the ability of ShaEP to retrieve active compounds. DUD has recently been found to contain structurally analogous ligand molecules when the ligands were clustered according to their reduced graph representations.[63] Sets of less analogous ligands for 13 targets extracted from the Wombat database have been made available at the DUD Web site.[68] Of these targets 11 are in the original DUD and therefore have a decoy set available. We used these 11 sets in place of the original DUD ligand sets in order to reduce the analogy bias.

The structures in the 11 sets of ligands from the Wombat database had 2D geometry, no hydrogen atoms, neutral form, and no stereochemistry indicators. Therefore, the ligand structures were processed using the ZINC (version 8) upload functionality,[69] which assigns formal charges and stereochemistry and generates 3D coordinates for the structures. The ZINC upload service was the closest match to the process applied to the decoy compounds[65] that we had access to. Balloon was then used to generate conformer ensembles for the ligands and the decoys using the same settings as listed above for the Jain data set. A total of 27 decoy structures, listed in the Supporting Information, were left out because they contained one or more atoms for which there is no MMFF94 atom type available, and therefore no partial atomic charges could be computed for these structures.

The cocrystallized ligand from the PDB file of each target[67] was used as the template structure for virtual screening. Ligand structures were downloaded from the DUD Web site, converted to the SD file format, and checked for atoms with inconsistent bonded valence and formal charge. ShaEP, like all superimposition methods that use MEP, is sensitive to the correct protonation state of the compounds. Therefore, the structures were processed with the ZINC upload service, and the protonation states of the processed structures were manually transferred back to the X-ray conformations except for the template molecule for the target ACE for which the protonated state suggested by the ZINC service was deemed improbable and the unprotonated form, as provided by DUD, was used. MMFF94-like partial atomic charges were assigned to the models using Balloon.

The program ROCS (Rapid Overlay of Chemical Structures)[8,70] was used as a reference virtual screening method on the DUD data sets. ROCS performs molecular alignment by gradient optimization of volume overlap computed using spherical Gaussians of two kinds, one set representing the shape-density for non-hydrogen atoms of the molecule and one set representing the chemical properties (acceptor, donor, hydrophobic, anion, cation, and ring atoms). The latter set of Gaussians is called a "color force field". The four initial alignments used in ROCS superimpose the axes of the moments of inertia of the structures.[8] By default, ROCS ranks compounds by the sum of similarity scores for both sets of Gaussians, called the ComboScore. ROCS was run with the default settings.

In order to provide a "baseline" for comparison with the 3D methods, the ScreenMD[71] chemical fingerprint (CF) 2D similarity method was used to screen DUD. A single structure from the original DUD distribution SD files or the ZINC processed form of the ligands from the Wombat database was used for each compound in the 2D screen. The Tanimoto dissimilarity was used as the metric for ranking the results. The default settings of the screenmd program were used except that 1.0 was used as the threshold for printing out dissimilarity values. Based on the previous results by Cleves and Jain[24] and Nicholls,[57] a 2D fingerprint method can be expected to perform well on the DUD data sets.

Table 1. Areas under the ROC Curves for the Jain Data Sets[a]

target	ShaEP AUC	*onlyshape* AUC
serotonin	0.91 ± 0.04	0.47 ± 0.05
Bzr	0.85 ± 0.06	0.84 ± 0.06
muscarinic	0.94 ± 0.02	0.57 ± 0.05
histamine	0.95 ± 0.02	0.83 ± 0.04

[a] The *onlyshape* column lists the results for using only volume overlap optimization instead of the full ShaEP algorithm.

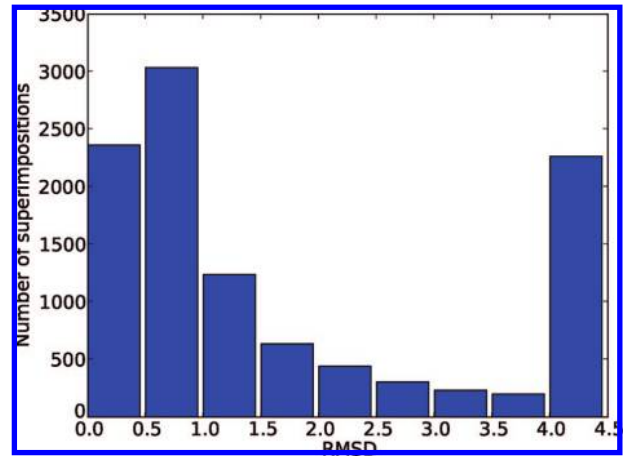

Figure 2. The distribution of the rmsd values ($N = 10718$) of all the generated superimpositions of X-ray ligand structures.

The area under a receiver-operating characteristic curve (AUC) was used as the performance metric of the programs in the virtual screening benchmark. A receiver-operating characteristic (ROC) curve depicts the fraction of true positives on the *y*-axis versus the fraction of false positivies on the *x*-axis found in a classification experiment. A random classification would produce a diagonal line (AUC = 0.5), and a perfect screening tool would produce a curve that runs from (0,0) via (0,1) to (1,1) (AUC = 1.0). AUC equals the probability of ranking a randomly selected active compound higher than a randomly selected inactive compound.[72] The ROC and AUC in this study were computed using the algorithms given by Fawcett.[73] The method of Hanley et al.[72] was used to calculate a conservative estimate of the standard error of AUC.

3. RESULTS

3.1. Reproduction of X-ray Superimpositions. A total of 10718 superimpositions were generated in the "sanity check" on the 542 ligand structures extracted from X-ray crystal structure models. Table 1 in the Supporting Information lists the average rmsd values for all ligands and the DPI value for those ligands for which the required information was available in the PDB file. Figure 2 depicts the distribution of the number of superimpositions within rmsd intervals. The number of superimpositions for which rmsd exceeds the

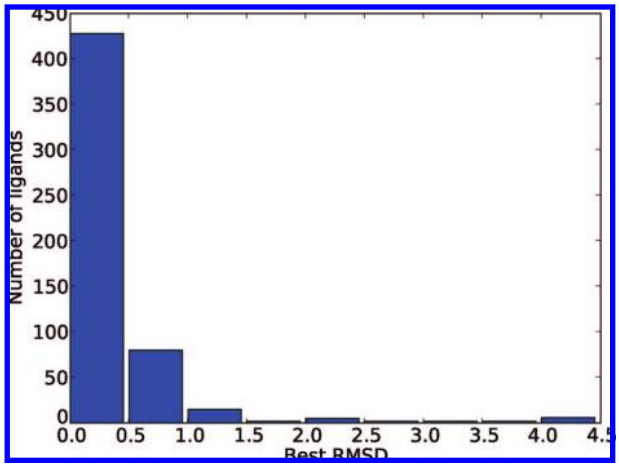

Figure 3. The distribution of the lowest rmsd values for each ligand ($N = 542$).

$2\sqrt{2\sigma}$ threshold was 1685 (49.4%) out of the 3414 for which the DPI could be computed. The number of wrong overlays was 3445 out of the 10718 superimpositions generated (32.1%), which indicates that the selected sets of ligands were not trivially easy to superimpose. The distribution of the average of the *lowest* rmsd obtained for each molecule, depicted in Figure 3 (with an average of 0.48 ± 0.93 Å), indicates that for nearly all of the ligands there exists at least one template structure that induces the correct orientation. There were 17 ligands that did not have a single correct superimposition.

The aim of the whole field-graph matching step in the ShaEP algorithm was to provide more reasonable starting points for the optimization of the overlap volume than a mere alignment according to shape (the four orientations obtained from the alignment of the principal axes of geometry). These additional starting orientations are meant to lead to the correct superimposition of molecules of different size yet having similar substructures. One example of such a pair of structures are the ligands in the PDB files 1e66 ((−)-huprine X) and 1q84 (anti-TZ2PA6); both ligands are inhibitors of acetylcholinesterase. The full ShaEP algorithm overlaid (−)-huprine X on anti-TZ2PA6 with an rmsd of 0.63 Å (Figure 4). If the initial orientations produced by the field-graph matching are not used (option *onlyshape*), the rmsd of the produced superimposition was 14.12 Å. It is evident that the field-graph matching step is necessary in order to reproduce the crystal structure orientation of (−)-huprine X with respect to anti-TZ2PA6.

3.2. Virtual Screening. The ROC curves for the virtual screening of the Jain data sets are presented in Figure 5. The areas under the curves (AUC) are listed in Table 1. In comparison with the results obtained by Cleves and Jain (Table 1 in ref 24), ShaEP and Surflex-Sim[64] produce equal AUC values within the error limits. Table 1 lists the results for running ShaEP with the *onlyshape* option on. Given the error marginals, the full ShaEP algorithm performs better than its shape-overlap-only version in three out of the four cases.

Table 2 lists the results for running ShaEP and ROCS on the DUD data sets. The average AUC for ShaEP was 0.64 ± 0.17, with a median of 0.64. The average AUC for ROCS was 0.69 ± 0.17 and median 0.68. Given the error estimates, ShaEP produced a higher AUC than ROCS in 6 out of the

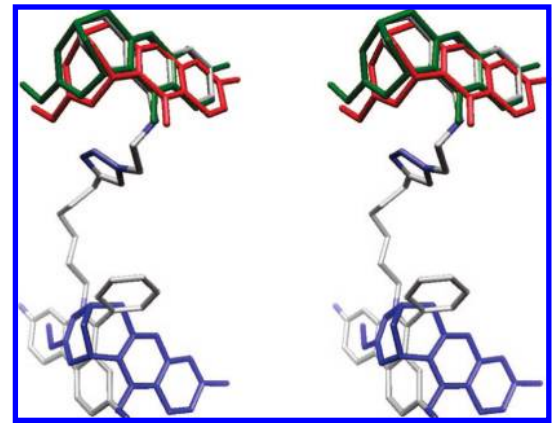

Figure 4. A wall-eyed stereoview of the model of (−)-huprine X as extracted from the PDB entry 1e66 (red), superimposed on anti-TZ2PA6 (the ligand in 1q84; colored according to elements), using the ShaEP algorithm (green) and using only volume overlap optimization with four initial orientations (blue). The figure was created using Bodil.[59]

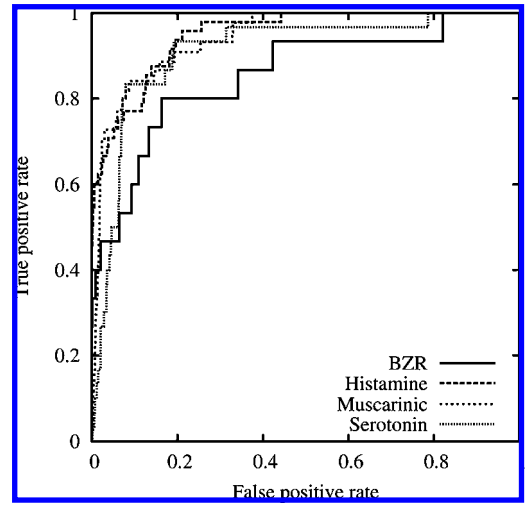

Figure 5. ROC curves for the virtual screening against four different targets of Jain[64] using the ZINC druglike decoy set.

40 cases, and ROCS had a higher AUC in 16 cases (18 ties between the 3D methods). The CF fingerprints gave the best AUC in 9 cases, with an average AUC of 0.72 ± 0.17 and a median of 0.73. Between the 3D methods, the processing times compare favorably for ROCS, which is almost an order of magnitude faster than ShaEP in some cases. For ShaEP, the average processing time per structure was 148 ± 90 ms with a median of 125 ms. For ROCS, the average time was 24 ± 5 ms with a median of 23 ms per structure. All computations were made on a 2.4 GHz Intel Xeon CPU (specification number SL687) running Linux.

Cheeseright et al.[74] used a subset of DUD for the evaluation of FieldScreen, a virtual screening program that uses a field-based method described earlier.[14] Comparison of ShaEP, ROCS, and FieldScreen can be made on a qualitative basis even if the differences in data preparation and the lack of error estimates on the results of Cheeseright et al., reproduced in Table 3, do not allow for exact ranking of the methods. It seems that none of these methods performs consistently with higher accuracy than the others. Execution times for FieldScreen were not published.

Table 2. Areas under the ROC Curves for the DUD Data Sets[67a]

target	ShaEP AUC	ROCS AUC	CF AUC	ShaEP ms/conf.	ROCS ms/conf.	no. of ligands/no. of decoys
ACE	0.58 ± 0.04	0.76 ± 0.04	0.88 ± 0.03*	134	24	49/1727
AChE	0.77 ± 0.03	0.75 ± 0.03	0.66 ± 0.03	151	30	105/3712
ADA	0.69 ± 0.06	0.85 ± 0.05	0.88 ± 0.04	74	18	23/821
ALR2	0.56 ± 0.04	0.52 ± 0.04	0.58 ± 0.04	75	17	51/918
AmpC	0.78 ± 0.06	0.86 ± 0.05	0.87 ± 0.05	58	17	21/732
AR	0.46 ± 0.04	0.48 ± 0.04	0.54 ± 0.04	95	21	62/2628
CDK2	0.46 ± 0.02	0.61 ± 0.02	0.65 ± 0.02	81	22	201/1778
COMT	0.34 ± 0.07	0.39 ± 0.08	0.57 ± 0.09*	89	15	11/430
COX-1	0.48 ± 0.06	0.54 ± 0.06	0.78 ± 0.06*	53	19	25/849
COX-2	0.61 ± 0.03	0.68 ± 0.03*	0.47 ± 0.03	135	22	99/12462
DHFR	0.56 ± 0.02	0.95 ± 0.01	0.94 ± 0.01	123	23	201/2126
EGFr	0.37 ± 0.03	0.62 ± 0.03*	0.50 ± 0.03	124	22	100/14893
$ER_{agonist}$	0.89 ± 0.03	0.92 ± 0.02	0.83 ± 0.03	100	20	67/2353
$ER_{antagonist}$	0.53 ± 0.03	0.78 ± 0.03	0.74 ± 0.03	226	32	108/1395
FGFr1	0.23 ± 0.02	0.57 ± 0.03	0.54 ± 0.03	133	27	118/4204
FXa	0.76 ± 0.02*	0.52 ± 0.02	0.47 ± 0.02	224	30	167/5092
GART	0.86 ± 0.05	0.79 ± 0.06	0.86 ± 0.04	227	31	21/753
GPB	0.86 ± 0.03	0.94 ± 0.02*	0.76 ± 0.04	82	18	52/1844
GR	0.64 ± 0.03	0.73 ± 0.03	0.85 ± 0.03*	183	26	78/2796
HIVPR	0.62 ± 0.04	0.55 ± 0.04	0.57 ± 0.04	551	35	53/1885
HIVRT	0.71 ± 0.03*	0.62 ± 0.03	0.61 ± 0.03	100	23	108/1437
HMGR	0.86 ± 0.04	0.78 ± 0.05	0.94 ± 0.03*	197	29	35/1241
HSP90	0.71 ± 0.06	0.80 ± 0.05	0.59 ± 0.05	152	24	24/860
InhA	0.71 ± 0.03	0.69 ± 0.03	0.75 ± 0.03	107	24	85/3035
MR	0.83 ± 0.07	0.82 ± 0.07	0.90 ± 0.05	154	26	15/535
NA	0.86 ± 0.03	0.98 ± 0.01*	0.80 ± 0.04	116	25	49/1743
P38 MAP	0.65 ± 0.04*	0.43 ± 0.03	0.43 ± 0.03	230	24	64/8387
PARP	0.57 ± 0.05	0.62 ± 0.05	0.85 ± 0.04*	62	18	33/1173
PDE5	0.63 ± 0.02*	0.52 ± 0.02	0.50 ± 0.02	219	29	155/1808
PDGFrb	0.30 ± 0.02	0.40 ± 0.02	0.66 ± 0.02*	122	24	157/5612
PNP	0.78 ± 0.05	0.90 ± 0.04	0.97 ± 0.02*	80	17	25/882
PPARg	0.74 ± 0.04*	0.65 ± 0.04	0.59 ± 0.04	282	36	52/2906
PR	0.58 ± 0.06	0.53 ± 0.06	0.61 ± 0.06	263	24	27/967
RXRa	0.90 ± 0.05	0.98 ± 0.02	0.97 ± 0.03	215	33	20/708
SAHH	0.76 ± 0.05	0.99 ± 0.01	0.98 ± 0.02	59	17	33/1156
SRC	0.48 ± 0.02	0.47 ± 0.02	0.84 ± 0.02*	128	26	155/5792
thrombin	0.71 ± 0.04	0.57 ± 0.04	0.72 ± 0.03	251	31	65/2292
TK	0.85 ± 0.05	0.86 ± 0.05	0.89 ± 0.05	58	16	22/784
trypsin	0.56 ± 0.05	0.69 ± 0.05	0.65 ± 0.04	137	31	44/1544
VEGFr2	0.49 ± 0.03	0.50 ± 0.03	0.46 ± 0.03	85	25	74/2641

[a] The highest AUC for each target, given the error limits, is indicated with an asterisk. The processing times for ShaEP are averaged over all structures in each set. For ROCS, the average times were computed over the decoy structures only. The ligand and decoy counts are based on structures with different names.

Table 3. AUC Values Obtained for a Subset of DUD Using Different Virtual Screening Methods[a]

target	ShaEP	ROCS	CF	FieldScreen
ACE	0.58	0.76	0.88	0.67
AChE	0.77	0.75	0.66	0.76
CDK2	0.46	0.61	0.65	0.47
COX-2	0.61	0.68	0.47	0.92
EGFr	0.37	0.62	0.50	0.84
FXa	0.76	0.52	0.47	0.74
HIVRT	0.71	0.62	0.61	0.70
InhA	0.71	0.69	0.75	0.71
P38 MAP	0.65	0.43	0.43	0.33
PDE5	0.63	0.52	0.50	0.66
PDGFrb	0.30	0.40	0.66	0.29
SRC	0.48	0.47	0.84	0.45
VEGFr2	0.49	0.50	0.46	0.48

[a] The AUC values for FieldScreen are reproduced from Table S10 (Supporting Information) in ref 74, For ShaEP, ROCS, and CF the results are obtained in this study.

4. DISCUSSION

The molecular alignments obtained using the field-graph method are often dominated by a single strongly positive or negative node when the overlap of a pair of vertices was scored according to the product of the ESP values at the vertex locations.[25,75] The alignments obtained by ShaEP should be less dominated by a single pair of matching vertices than alignments obtained using other field-graph methods because the exponential weighting term in eq 7 rewards the overlap of vertices with *similar* potential without reference to its magnitude.

The results of the virtual screening experiment on the Jain data set indicate that inclusion of the field-graph information improves the recognition of active compounds. Still, the field-graph method is known to miss about 6% of structures that are known to be similar to the template structure.[3] In order to alleviate this, ShaEP uses four starting orientations that superimpose the geometric centroids of the molecules and aligns the principal axes of geometry in addition to the transformations obtained from the matching of the field-graphs. Because of these "fall-back" initial orientations, one can expect ShaEP to do better than a purely field-graph based method in terms of missed actives. Some of the results presented in Table 3, e.g. those for COX-2, suggest that

adding vertices at the locations of local extrema of the MEP,[14] as FieldScreen does, might provide additional accuracy. This option is not presently implemented in ShaEP.

The template structures used for some biological targets in the DUD benchmark are obviously inappropriate because both ShaEP and ROCS produced AUC values close to 0.5, which would be obtained by a random ranking of the compounds. Some AUC values are even closer to zero than to 0.5, which indicates that the template is better at picking inactive molecules than active ones. The logic of assigning compounds as actives could be inverted in these cases if the behavior would be known *a priori*, which obviously does not apply to real-life situations. A publicly available set of reasonable template structures would ameliorate the use of DUD as a benchmark for ligand based virtual screening methods, although the data set is still plagued with analogy bias as indicated by the high mean AUC for the 2D fingerprint method.

The quality of the partial atomic charges assigned to the structures is essential for the reproduction of the MEP. The MMFF94-like charges used in this study may not be the best alternative because they do not depend on the molecular conformation. Even if the MEP could be reproduced with high accuracy ShaEP is still sensitive to the correct protonation state of the screening compounds, which becomes particularly evident in the cases of FGFr1, DHFR, and ACE data sets that show a low AUC for ShaEP and relatively high AUC for ROCS. The template structure for the FGFr1 set is neutral, and 79 of the 118 ligands are charged. The ACE template is a zwitterion that contains a protonated primary amine and two deprotonated carboxylic acids, whereas only 12 of the 49 ligands are zwitterions. The template for the DHFR set contains two deprotonated carboxylic groups, but 18 of the 201 ligands are double deprotonated (of which 9 are zwitterions also carrying a positive charge). Moffat et al. observed a similar effect of charge on the screening results using a field-graph method.[51] Cheeseright et al. used dampened formal charges in the computation of electrostatic potential energy in order to avoid "field distortions".[14] These observations suggest that neutral compounds should be used in screening and/or that the relative weight w in the final score (eq 9) should be adjusted to give less emphasis to the electrostatic similarity in the final score value. It can be even argued that the similarity of MEPs is not a suitable basis for assessing the similarity of ligands because of the intrinsic inability to account for the polarization effects due to a host protein. The color force field used in ROCS or some other scoring scheme that reflects the local chemical character of the structure is therefore a viable alternative to the MEP based score.

According to the timing results in Table 2, ShaEP is clearly slower than ROCS. The use of a graph-matching step, a set of Gaussians to represent the field-graph vertices, and more than four initial orientations for the optimization of the overlay impose a higher computational cost on ShaEP. The use of these steps is appropriate in situations where the template and the screened structures differ in size. The evaluation of the volume overlap using eq 6 accounts for a large portion of the total computing time. Approximations to the overlap integral expression that preserve the location of the maximum of the function would gain speed without affecting the outcome, because the final similarity score would still be the same. Even without this optimization, ShaEP can process large virtual libraries within a reasonable time frame, for example, in the median case (125 ms per structure) 10^6 structures in less than two days.

Our future work will concentrate on improved ways of scoring the alignments. The effect of the partial charges on the screening results and MEP reproduction is a subject of ongoing research in our laboratory.

5. CONCLUSIONS

Molecular similarity analysis is a widely studied topic because of its practical value to drug design. While there exists a wide array of software tools for this purpose, each tool has its shortcomings. In this study, we implemented a molecular alignment/similarity analysis program that combines elements of previously published algorithms in order to overcome some of their limitations. In particular, the program ShaEP introduced in this study is expected to find reasonable molecular alignments even for ligands of different size, which may be of importance in the preparation for a 3D-QSAR study or in the virtual screening of a database of reagents against a large drug. The implemented method was validated against several test cases of small ligands that bind to a biological target. The execution speed allows for the virtual screening of large structural databases.

ACKNOWLEDGMENT

We thank Dr. Susanna Repo for her critical reading of this manuscript and Mikko Huhtala for the help with Contactos. The Academy of Finland, Sigrid Jusélius Foundation, Magnus Ehrnrooth Foundation, the Tor, Joe and Pentti Borg Foundation, and Stiftelsens för Åbo Akademi forskningsinstitut are acknowledged for their financial support. The Structural Bioinformatics Laboratory belongs to the Center of Excellence in Cell Stress of Åbo Akademi University.

Supporting Information Available: PDB identifiers of the structure files used in the superimposition test runs and the names of compounds excluded from the virtual screening experiment. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615–633.

(2) Bemis, G. W.; Kuntz, I. D. A Fast and Efficient Method for 2D and 3D Molecular Shape Description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607–628.

(3) Wild, D.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 159–167.

(4) Thorner, D.; Wild, D.; Willett, P.; Wright, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900–908.

(5) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80–86.

(6) Watson, P.; Willett, P.; Gillet, V. J.; Verdonk, M. L. Calculating the Knowledge-Based Similarity of Functional Groups Using Crystallographic Data. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 835–857.

(7) Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. FLUFF-BALL, A Template-Based Grid-Independent Superposition and QSAR Technique: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Model* **2003**, *43*, 1780–1793.
(8) Rush, T.; Grant, J.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.
(9) Labute, P.; Williams, C.; Feher, M.; Sourial, E.; Schmidt, J. Flexible Alignment of Small Molecules. *J. Med. Chem.* **2001**, *44*, 1483–1490.
(10) Pitman, M. C.; Huber, W. K.; Horn, H.; Krämer, A.; Rice, J. E.; Swope, W. C. FLASHFLOOD: A 3D Field-Based Similarity Search and Alignment Method for Flexible Molecules. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 587–612.
(11) Putta, S.; Landrum, G.; Penzotti, J. Conformation Mining: An Algorithm for Finding Biologically Relevant Conformations. *J. Med. Chem.* **2005**, *48*, 3313–3318.
(12) Tervo, A.; Ronkko, T.; Nyrönen, T.; Poso, A. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. 1. Alignment and Virtual Screening Applications. *J. Med. Chem.* **2005**, *48*, 4076–4086.
(13) Rönkkö, T.; Tervo, A. J.; Parkkinen, J.; Poso, A. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. II. Description and Characterization. *J. Comput.-Aided Mol. Des* **2006**, *20*, 227–236.
(14) Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model.* **2006**, *46*, 665–676.
(15) Good, A. C. Novel DOCK Clique Driven 3D Similarity Database Search Tools for Molecule Shape Matching and beyond: Adding Flexibility to the Search for Ligand Kin. *J. Mol. Graphics Modell.* **2007**, *26*, 656–666.
(16) Cho, S.; Sun, Y. FLAME: A Program to Flexibly Align Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 298–306.
(17) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.
(18) Mavridis, L.; Hudson, B.; Ritchie, D. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *J. Chem. Inf. Model.* **2007**, *47*, 1787–1796.
(19) Marin, R.; Aguirre, N.; Daza, E. Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials. *J. Chem. Inf. Model.* **2008**, *48*, 109–118.
(20) Good, A. C.; Richards, W. G. Excplicit Calculation of 3D Molecular Similarity. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 321–338.
(21) Lemmen, C.; Lengauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215–232.
(22) Maldonado, A.; Doucet, J.; Petitjean, M.; Fan, B.-T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Diversity* **2006**, *10*, 39–79.
(23) Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field Interaction and Geometrical Overlap: A New Simplex and Experimental Design Based Computational Procedure for Superposing Small Ligand Molecules. *J. Med. Chem.* **2003**, *46*, 1359–1371.
(24) Cleves, A.; Jain, A. Effects of Inductive Bias on Computational Evaluations of Ligand-Based Modeling and on Drug Discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 147–159.
(25) Thorner, D. A.; Willet, P.; Wrigth, P. M.; Taylor, R. Similarity Searching in Files of Three-Dimensional Chemical Structures: Representation and Searching of Molecular Electrostatic Potentials Using Field-Graphs. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 163–174.
(26) Xian, B.; Li, T.; Sun, G.; Cao, T. The Combination of Principal Component Analysis, Genetic Algorithm and Tabu Search in 3D Molecular Similarity. *J. Mol. Struct.: THEOCHEM* **2004**, *674*, 87–97.
(27) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503–3510.
(28) 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**, *17*, 1653–1666.
(29) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188–191.
(30) Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112–116.
(31) *PyMol, version 1.1beta1*; DeLano Scientific LLC: Palo Alto, CA, U.S.A., 2008.
(32) Krissinel, E. B.; Henrick, K. Common Subgraph Isomorphism Detection by Backtracking Search. *Software Pract. Exper.* **2004**, *34*, 591–607.
(33) Kearsley, S. K. Structural Comparisons Using Restrained Inhomogeneous Transformations. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1989**, *45*, 628–635.
(34) Tomita, E.; Kameda, T. An Efficient Branch-and-bound Algorithm for Finding a Maximum Clique with Computational Experiments. *J. Glob. Optim.* **2007**, *37*, 95–111.
(35) Konc, J.; Janezic, D. An Improved Branch and Bound Algorithm for the Maximum Clique Problem. *MATCH Commun. Math. Comput. Chem.* **2007**, *58*, 569–590.
(36) Hofbauer, C.; Lohninger, H.; Aszodi, A. SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 837–847.
(37) Zefran, M.; Kumar, V.; Croke, C. Metrics and Connections for Rigid-Body Kinematics. *Int. J. Robot. Res* **1999**, *18*, 242−1–24216.
(38) Kavan, L.; Collins, S.; O'Sullivan, C.; Zara, J. *Dual Quaternions for Rigid Transformation Blending*; Technical report; Trinity College Dublin, 2006.
(39) Karney, C. F. Quaternions in Molecular Modeling. *J. Mol. Graphics Modell.* **2007**, *25*, 595–604.
(40) Clifford, W. K. *Mathematical Papers*; Macmillan & Co.: London, 1882.
(41) Park, S. I.; Shin, H. J.; Shin, S. Y. On-Line Locomotion Generation Based On Motion Blending. In *SCA '02: Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*; New York, NY, U.S.A. ACM: 2002; pp 105−111.
(42) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502–4520.
(43) Griewank, A. O.; Markey, B. R.; Evans, D. J. Singularity-Free Static Lattice Energy Minimization. *J. Chem. Phys.* **1979**, *71*, 3449–3454.
(44) Schlick, T.; Fogelson, A. TNPACK - A Truncated Newton Minimization Package for Large-Scale Problems: I. Algorithm and Usage. *ACM Trans. Math. Softw.* **1992**, *18*, 46–70.
(45) Schlick, T.; Fogelson, A. TNPACK -- A Truncated Newton Minimization Package for Large-Scale Problems: II. Implementation Examples. *ACM Trans. Math. Softw.* **1992**, *18*, 71–111.
(46) Xie, D.; Schlick, T. Efficient Implementation of the Truncated-Newton Algorithm for Large-Scale Chemistry Applications. *SIAM J. Optim.* **1999**, *10*, 132–154.
(47) Xie, D.; Schlick, T. Remark on Algorithm 702-the Updated Truncated Newton Minimization Package. *ACM Trans. Math. Softw.* **1999**, *25*, 108–122.
(48) Cleves, A.; Jain, A. Robust Ligand-Based Modeling of the Biological Targets of Known Drugs. *J. Med. Chem.* **2006**, *49*, 2921–2938.
(49) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quantum Chem.* **1987**, *32*, 105–110.
(50) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(51) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A Comparison of Field-Based Similarity Searching Methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model* **2008**, *48*, 719–729.
(52) Cruickshank, D. W. J. Remarks about Protein Structure Precision. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 583–601.
(53) Blow, D. M. Rearrangement of Cruickshank's Formulae for the Diffraction-Component Precision Index. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 792–797.
(54) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: Pharmacophore-Based Protein-Ligand Docking. *J. Med. Chem.* **2004**, *47*, 6804–6811.
(55) Goto, J.; Kataoka, R.; Muta, H.; Hirayama, N. ASEDock -- Docking Based on Alpha Spheres and Excluded Volumes. *J. Chem. Inf. Model.* **2008**, *48*, 583–590.
(56) Hawkins, P.; Warren, G.; Skillman, A.; Nicholls, A. How to Do an Evaluation: Pitfalls and Traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.
(57) Nicholls, A. What Do We Know and When Do We Know It. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
(58) Lehtonen, J. V.; Vainio, M. J.; Hoffrén, A.-M.; Johnson, M. S. VERTAA: Protein Superimposition Based on C-α Packing Density Profiles. manuscript in preparation, 2008.
(59) Lehtonen, J. V.; Still, D.-J.; Rantanen, V.-V.; Ekholm, J.; Björklund, D.; Iftikhar, Z.; Huhtala, M.; Repo, S.; Jussila, A.; Jaakkola, J.; Pentikäinen, O.; Nyrönen, T.; Salminen, T.; Gyllenberg, M.; Johnson, M. S. BODIL: A Molecular Modeling Environment for Structure-Function Analysis and Drug Design. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 401–419.
(60) Vainio, M.; Johnson, M. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.
(61) *Maestro, version 8.0*; Schrödinger, LLC: New York, NY, 2007.
(62) Huhtala, M. *Contactos, version 1.1.9*. http://www.abo.fi/∼ mhuhtala/contactos.html (accessed March 28, 2008).

(63) Good, A.; Oprea, T. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.
(64) Jain, A. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47*, 947–961.
(65) Irwin, J.; Shoichet, B. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(66) Pham, T.; Jain, A. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856–5868.
(67) Huang, N.; Shoichet, B.; Irwin, J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(68) http://dud.docking.org/ (accessed May 8, 2008).
(69) http://zinc.docking.org/ (accessed May 8, 2008).
(70) *ROCS, version 2.3.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2008.
(71) *ScreenMD, version 5.1.2*; ChemAxon Kft.: Budapest, Hungary, 2008.
(72) Hanley, J.; McNeil, B. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29–36.
(73) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
(74) Cheeseright, T.; Mackey, M.; Melville, J.; Vinter, J. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *J. Chem. Inf. Model.* **2008**, *48*, 2108–2117.
(75) Miller, M.; Sheridan, R.; Kearsley, S. SQ: A Program for Rapidly Producing Pharmacophorically Relevent Molecular Superpositions. *J. Med. Chem.* **1999**, *42*, 1505–1514.

CI800315D